Journal of Computers Vol. 34 No. 5, October 2023, pp. 15-27
doi: 10.53106/199115992023103405002

Support Vector Machine Based Automatic Classification Method for IoT Big Data Features

Yong-Hua Xu[1,2*]

[1] School of Computer Engineering, Jinling Institute of Technology,
Nanjing, 211169, China
[2] Jiangsu Key Laboratory of Data Science and Smart Software, Jinling Institute of Technology,
Nanjing, 211169, China
xuyhua0730@163.com

Received 30 August 2022; Revised 30 December 2022; Accepted 10 February 2023

Abstract. As China's information technology development shifts from a single high-speed growth stage to a multidimensional high-quality development stage, the Internet of Things (IoT) enters all aspects of life and becomes more and more popular. The demand for IoT big data information analysis and processing is increasing, and the important role of feature automatic classification methods becomes increasingly prominent. This research proposes SPO-SVM and WSPO-SVM models based on support vector machine for smart home environment monitoring data under the big data of Internet of Things, and then optimizes them with particle swarm optimization algorithm and adaptive method. Finally, the data set is selected for comparative experimental analysis of each optimization algorithm model. The experimental results show that the optimized WSPO-SVM model has less total misclassification and single class misclassification compared with other algorithms under Wine dataset. In cross-validation, both its classification accuracy performance outperformed other algorithms. Under 10 sets of smart home environment monitoring data sets, the WSPO-SVM algorithm model achieves 100% accuracy in 6 out of 10 test data sets, with an average accuracy of 97.67%, which is about 9% higher than the ordinary SVM algorithm model and about 15% higher than other feature classification algorithms. The experimental results prove that the WSPO-SVM algorithm can complete the feature classification work in the IoT big data environment, which meets the expectation.

Keywords: internet of things, SVM, SPO, feature classification algorithm

1 Introduction

With the rapid development of Internet technology, the amount of global information data has shown explosive growth. In the face of massive information data, the biggest challenge we face is how to extract valuable information from massive data [1]. In recent years, data mining technology has been widely used in e-commerce, medical insurance, financial securities and other fields [2]. With the advent of the big data era of the Internet of Things, the value of data feature classification begins to appear, which has a significant impact on social progress, economic development and people's lives [3]. All walks of life have begun to attach importance to the exploration of the Internet of Things. Data processing has gradually become an indispensable part of the development of the Internet of Things, which has had a positive impact on our life, work and way of thinking [4]. Research on big data of the Internet of Things is not only a powerful tool to improve people's living standards and promote industrial transformation and upgrading, but also an important means to promote scientific, intelligent and personalized services for various life related appliances [5].

At present, there is a general lack of research on automatic classification methods of data features in various fields of the Internet of Things, and there is also a lack of comprehensive analysis of multi-dimensional data related to smart devices and real life values, especially smart home neighborhoods. With the technological progress of the Internet of Things, the concept of smart home has developed rapidly in recent years. However, at present, the data mining ability of the Internet of Things of smart home is very weak, the information processing efficiency is low, and the classification is not accurate enough. In order to improve the information processing efficiency of the Internet of Things and meet the classification needs of data features in the big data environment, the current research should not only focus on the improvement of hardware level, but also effectively improve the data

* Corresponding Author

mining methods.

In order to improve the data mining ability of the Internet of Things, this study proposes a support vector machine classification model based on adaptive optimization and particle swarm optimization algorithm to solve the problems of insufficient accuracy and low classification efficiency of the current smart home Internet of Things data classification. The environmental monitoring part of smart home is discussed in depth, and the proposed method is used to upgrade the environmental monitoring system of smart home, which improves the accuracy and accuracy of environmental monitoring. At the same time, the research is not limited to environmental monitoring. Through tests on different data sets, it is proved that the proposed method is applicable to all kinds of Internet of Things data mining and data classification scenarios, which has contributed some strength to the development of the Internet of Things.

The research content mainly includes four parts. The second part is an overview of the research status of automatic classification methods of IoT big data features at home and abroad; In the third part, an automatic classification method of IoT big data features based on support vector machine is proposed. In the first section, a classification model of support vector machine is established. In the second section, a SVM model based on PSO and adaptive optimization is constructed. In the third section, data analysis is made on the adaptive smart home environment monitoring system; The fourth part verifies the application effect of the improved SVM Internet of Things big data feature automatic classification model. The results show that the automatic classification method of IoT big data features based on support vector machine has good application effect and application prospect.

2 Related Work

To promote the benign and sustainable development of IoT, the study of automatic classification methods for IoT big data features has become a focal topic nowadays, for which researchers at home and abroad have also conducted in-depth studies. Al-Thanoon et al considered the common problem of big data that many features may be irrelevant, and proposed the binary crow search algorithm (BCSA) for this problem. In order to improve the classification performance with reasonable feature selection, an improved method for determining flight length parameters based on comparative learning strategy and BCSA concept is proposed [6]. Y. Wang and his experimental partners proposed to design a new robust loss function based on minimizing the misclassification cost to deal with the imbalanced classification problem in a noisy environment, and they applied the proposed loss function to a support vector machine (SVM) [7]. Sevinc et al. proposed an evolutionary feature selection algorithm combined with single hidden layer feedforward neural network (SLFN), which finds the most efficient feature subset and provides the best prediction accuracy. This study combined with the evolutionary technique of genetic algorithm (GA), and used extreme learning machine (ELM) to calculate the adaptation value for each selected feature subset (prediction accuracy) [8]. Tang et al. argued that in classifying large-scale datasets, marking a sufficient number of training samples is both time-consuming and laborious, and it is difficult to train models in a time-efficient and high-precision manner. For large-scale datasets, there is a trade-off between speed and accuracy. Aiming at this problem, a new classification strategy for large-scale data was proposed by combining the K-mean clustering technique with a multicore support vector machine approach [9]. Gaye and his team introduced fuzzy affiliation into multicore learning and found that the time complexity of the primal problem is determined by the number of dimensions. And the time complexity of the dyadic problem is determined by the number, which constitutes the scale of the data. Transforming the dyadic solution of the problem into a classification surface in the primal space can increase the speed of the algorithm and can make traditional machine learning algorithms meet the requirements of the big data era [10].

Janani and other researchers proposed proposed an artificial bee colony (ABCFS) based feature selection algorithm to improve the accuracy of classification [11]. Wei and his team concluded that for unbalanced datasets, the traditional fuzzy support vector machine (FSVM) algorithm is not effective in classification and the introduced parameters are not optimized. In order to solve this problem, an improved fuzzy support vector machine algorithm, PSODECIFSVM, is proposed based on the particle swarm optimization algorithm. The algorithm designs the compactness of the fuzzy membership function and the information content of the sample according to the distance from the training sample to the center. Then, combining IFSVM algorithm with different error cost (DEC) algorithms, DECIFSVM algorithm is obtained. Finally, particle swarm optimization algorithm is used to optimize the parameters introduced in DECIFSVM algorithm [12]. Yin et al. proposed a new support vector machine parameter optimization method, which uses an improved whale-of algorithm (WOA) and an external archiving strategy, the advanced whale optimization algorithm (AWOA), to establish AWOA-

based SVM parameter optimization framework [13]. Saputra et al compared five algorithms in classification methods to obtain better performance in classification problems. The researchers analyzed and tested five algorithms classification using four different datasets as a tool for big data classification problems. The results of the study showed that in using the UCI knowledge base of 4 datasets to calculate AUC values, the SVM method outperformed the 4 comparative methods [14]. Kalita et al. researchers designed a new framework which uses new optimization modules Knowledge Based Search (KBS) and Moth-Flame Optimization (MFO) to perform optimization and efficiently train SVMs in a dynamic environment [15]. Firdausanti and his team wrapper based algorithms are Crazy Particle Swarm Optimization algorithm (CRAZYPSO) and Advanced Binary Ant Colony Optimization algorithm (ABACO) using k-fold cross validation accuracy to compare CRAZYPSO and ABACO algorithms for feature selection in the case of microarray data classification using support vector machine classification [16].

To sum up, it can be seen that there are still many technical drawbacks in the current automatic classification method of IoT big data features. For example, the classification methods in literature [6], literature [7] and literature [10] still have problems such as difficult data noise processing, low data processing accuracy, and poor data processing efficiency. Therefore, a SVM optimization algorithm based on PSO and adaptive optimization is proposed to study the automatic classification method of IoT big data features. Compared with the methods proposed in literature [12] and literature [13], the research has improved the data processing capability through more efficient and accurate adaptive methods. It can better solve the problems existing in the current data mining of the Internet of Things, and has a certain role in promoting the development of the data processing of the Internet of Things.

3 Research on Automatic Classification Method of IoT Big Data Features Based on Support Vector Machine

3.1 Research on Automatic Feature Classification Methods in the IoT Big Data Environment

With the development of the times, the Internet of Things has been popularized to conventional household appliances. This makes the previous single data processing mode no longer meet the needs of the times. This also puts forward new requirements for data feature classification methods [17]. To address this problem, this study uses the SVM algorithm to investigate the automatic classification method for IoT big data features. The core idea of the SVM algorithm is to map the data samples in high-dimensional space with a nonlinear map function, and then obtain the surface with a linear method in high-dimensional space, so as to separate the training sample points and maximize the distance between the training sample points and the optimal separation surface, and its architecture is shown in Fig. 1.

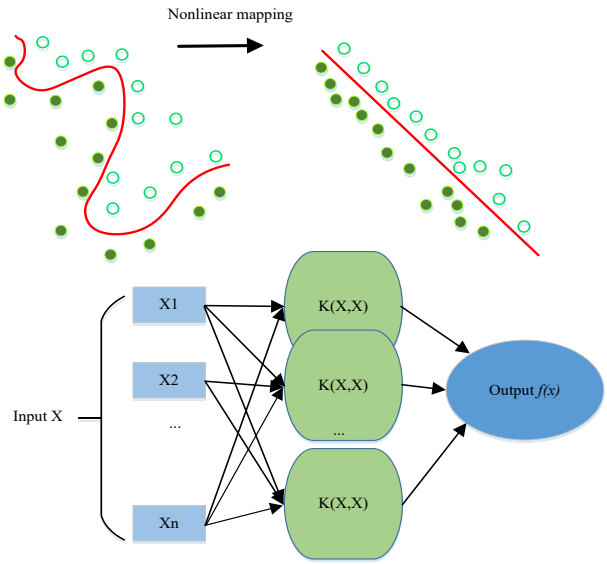

Fig. 1. Architecture of support vector machine

An example of linear data can be divided into an infinite number of levels, but with only one maximum classification interval. Let the training sample set $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, $y_i \in \{-1, 1\}$, $i = 1, 2, \ldots, n$. Then its optimal classification hyperplane is shown in Eq. (1).

$$\begin{cases} \min \frac{1}{2}\|w\|^2 \\ y_i\left(wx_i + b\right) - 1 \geq 0, i = 1, 2, ..., n \end{cases}. \tag{1}$$

In Eq. (1), w is the hyperplane normal vector, x_i is the feature of the i sample, y_i is the corresponding category label, and the minimum value of $||w||^2$ under the constraint is solved by using the Lagrange function method, which corresponds to the Lagrange function shown in Eq. (2).

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i\left(wx_i + b\right) - 1\right). \tag{2}$$

In Eq. (2), α_i is the Lagrange multiplier. The optimal conditions are obtained by solving the differential equations for b and w, respectively, as shown in Eq. (3).

$$\begin{cases} \nabla_b L\left(w, b, \alpha\right) = -\sum_{i=1}^{n} y_i \alpha_i = 0 \\ \nabla_w L\left(w, b, \alpha\right) = w - \sum_{i=1}^{n} y_i x_i \alpha_i = 0 \end{cases}. \tag{3}$$

The solution of Eq. (3) can be obtained by solving Eq. (4) in conjunction with Eq. (3).

$$\begin{cases} w = \sum_{i=1}^{n} y_i x_i \alpha_i \\ \sum_{i=1}^{n} y_i \alpha_i = 0 \end{cases}. \tag{4}$$

Bringing Eq. (4) to $L(w, b, \alpha)$ can get its corresponding pairwise form and constraints as shown in Eq. (5), and solving the problem, the optimization function can be obtained as shown in Eq. (6).

$$\begin{cases} \min_{\alpha} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \left(x_i \cdot x_j\right) - \sum_{i=1}^{n} \alpha_i \\ \sum_{i=1}^{n} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, ..., n \end{cases}. \tag{5}$$

$$f\left(x\right) = \operatorname{sgn}\left(\sum_{i=1}^{n} a_i^* y_i \left(x_i \cdot x\right) + b^*\right). \tag{6}$$

In the training example, there will be points that do not satisfy the condition, and by removing these points, the rest of the training remains linear. For the case where there are idiosyncratic points, the relaxation variable $\xi_i (\xi_i \geq 0)$ needs to be introduced, where the optimal classification hyperplane and constraints are shown in Eq. (7).

$$\begin{cases} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \\ y_i\left(wx_i + b\right) - 1 + \xi_i \geq 0, i = 1, 2, ..., n \end{cases}. \tag{7}$$

In Eq. (7), C is the penalty factor. When C is larger, it means that the penalty for misclassified samples is larger, and vice versa. The penalty factor allows the classification interval to be as large as possible while the misclassified samples are as small as possible. The problem is solved by the corresponding Lagrangian function method, as shown in Eq. (8).

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i(y_i(wx_i+b)-1+\zeta_i)-\sum_{i=1}^{n}\beta_i\xi_i\ . \tag{8}$$

In Eq. (8), α_i, β_i is the Lagrange multiplier, and the three optimal conditions shown in Eq. (9) are obtained by solving the differential equations for b, w and ξ_i, respectively.

$$\begin{cases}\nabla_b L(w,b,\alpha,\beta)=-\sum_{i=1}^{n}y_i\alpha_i=0\\ \nabla_w L(w,b,\alpha,\beta)=w-\sum_{i=1}^{n}y_i x_i\alpha_i=0\\ \nabla_{\xi_i} L(w,b,\alpha,\beta)=C-\alpha_i-\beta_i=0\end{cases}. \tag{9}$$

Solving the problem in the same way, the final optimization function is obtained as shown in Eq. (10).

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{n}a_i^{*}y_i(x_i\cdot x)+b^{*}\right). \tag{10}$$

Solving linear problems, the process is relatively simple, but the actual data are often nonlinear, so a nonlinear transformation is required, as shown in Eq. (11).

$$\phi: x\in R^n\rightarrow\phi(x)\in R^m, m>p\ . \tag{11}$$

Let the training sample set $D=\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, $y_i\in\{-1, 1\}$, $i=1, 2, \ldots, n$. Its optimal classification hyperplane is shown in Eq. (7), and the corresponding constraint becomes related to the mapping function $\phi(x_i)$ as shown in Eq. (12).

$$y_i(w^T\phi(x_i)+b)-1+\xi_i\geq 0,\ i=1, 2, ..., n. \tag{12}$$

Mapping data to a high-dimensional feature space requires the introduction of kernel functions in order to avoid the dimensional catastrophe problem. The use of kernel functions can effectively improve the algorithm's ability to handle linearly indistinguishable problems and make the classification model simpler. The common kernel functions are: linear kernel function, polynomial kernel function, and Gaussian kernel function. The Gaussian kernel function is used in this study, as shown in Eq. (13).

$$K(x,x_i)=\exp\left(-\gamma\|x-x_i\|^2\right),\gamma=\frac{1}{2\sigma^2}\ . \tag{13}$$

In Eq. (13), σ is the kernel width, which determines the magnitude of the correlation between the support vectors. x_i is the center of the kernel function, $\|x-x_i\|^2$ is the Euclidean distance between the vectors x and x_i, and the value of the Gaussian kernel function decreases monotonically as the spacing increases, at which point the optimal function is shown in Eq. (14).

$$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{n}a_i^{*}y_i\exp\left(-\frac{\|x-x_i\|^2}{2\sigma^2}\right)+b^{*}\right). \tag{14}$$

C σ The support vector machine classification algorithm has strong generality and is one of the common methods to effectively solve practical problems. The main parameters that affect the efficiency and accuracy of the support vector machine model are the penalty factor and the kernel parameter, which have the greatest impact on the classification results. They together determine the prediction ability of the model for unknown data, and seeking to optimize them will optimize the SVM algorithm.

3.2 SVM Algorithm Based on PSO and Adaptive Optimization

The particle swarm algorithm arose from imitating birds foraging for food, and the food that the birds are looking for is the optimal solution to the problem [18]. Let the space of the problem be in H dimension and the solution of the population l is the particles of the population. Each particle has three attributes, which are fitness value, position and velocity. The velocity and position of the particle i are denoted by V_i, $i\in(1, 2, \ldots, n)$, X_i, $i\in(1, 2, \ldots, n)$, the optimal position of the i particle passing through is P_i, $i\in(1, 2, \ldots, n)$, and the optimal position of the whole particle population passing through is G_i, $i\in(1, 2, \ldots, n)$. During the evolution of the PSO algorithm, the particles continuously optimize their positions and update their velocities, as shown in Eq. (15).

$$\begin{cases} v_{ij}^{k+1} = wv_{ij}^{k} + c_1 r_1 \left(P_j - x_{ij}^{k}\right) + c_2 r_2 \left(G_j - x_{ij}^{k}\right) \\ x_{ij}^{k+1} = x_{ij}^{k} + v_{ij}^{k+1} \end{cases}. \tag{15}$$

In Eq. (15), w denotes the inertia weight, c_1, c_2 denote the learning factor, r_1, r_2 are the random number in the interval of [0,1], and k denotes the evolutionary algebra. From the equation, it can be concluded that the update of particles is determined by three components: inertia, cognitive, and social. The inertia component is mainly expressed in the role of the current velocity, and the particle will follow the proportion of inertia weights to develop itself. The cognitive component is mainly expressed in the role of the individual optimal position, and the particle will evolve in the direction of the individual optimal position, showing the local search ability of the particle. The social component is mainly manifested in the role of group optimal position, and the particles will follow the direction of group evolution and evolve continuously. Based on the optimization of the SVM model by the PSO algorithm, the flow of the obtained PSO-SVM algorithm model is shown in Fig. 2.

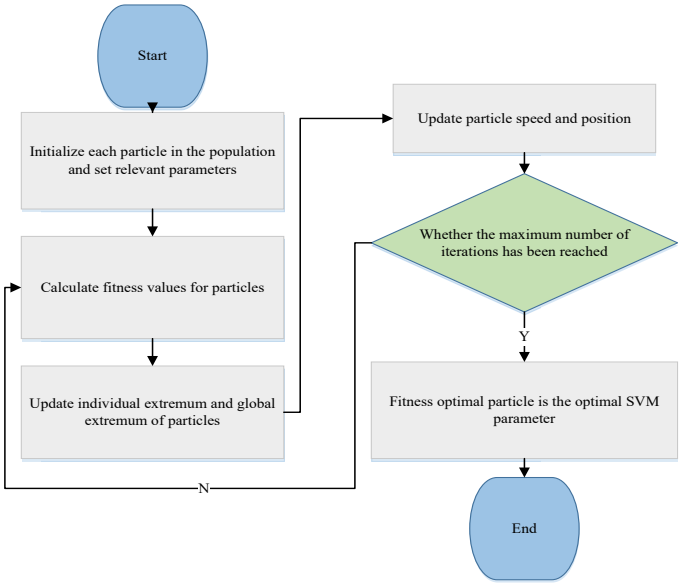

Fig. 2. Flow chart of PSO-SVM algorithm

Compared with the original SVM algorithm model, the optimized PSO-SVM algorithm model not only combines local search with population search, but also determines the local optimal position and population optimal position through the memory function of particles. This approach has a faster rate of evolution. However, this algorithm model cannot avoid the problem that PSO algorithm will easily fall into local optimal solution and keep high dependence on parameters. This study proposes to improve the particle swarm algorithm by using adaptive methods to address these problems. This method can get rid of the dilemma of falling into the local optimal solution and make the calculation result close to the global optimal solution. The adaptive weights are calculated as shown in Eq. (16).

$$w_i = w_{\max} - \frac{w_{\max} - w_{\min}}{1 + \exp\left(\left|\frac{f_i - f_{avg}}{f_g - f_{avg}}\right|\right)}. \tag{16}$$

In Eq. (16), $w_{\max}$, $w_{\min}$ represents the maximum and minimum values of the weights, usually the maximum value is 0.9 and the minimum value is 0.4. f_i represents the fitness of particle i, f_{avg} represents the average fitness of the population, and f_g represents the optimal fitness of the population. Then the update of the position and velocity corresponding to the i th particle is shown in Eq. (17).

$$\begin{cases} x_{ij}^{k+1} = x_{ij}^{k} + v_{ij}^{k+1} \\ v_{ij}^{k+1} = w_i v_{ij}^{k} + c_1 r_1 \left(P_j - x_{ij}^{k}\right) + c_2 r_2 \left(G_j - x_{ij}^{k}\right) \end{cases}. \tag{17}$$

In equation (17), c_1, c_2 is the learning factor, which is a constant, r_1, r_2 is a random number in the interval of [0,1], w_i is the adaptive weight of the particle, and k represents the number of generations of evolution. When there is a particle in the population that converges to the local optimal solution, other particles will also be influenced by it and converge to its local optimal solution, which will lead the population to finish convergence early and fall into the local optimal solution. To address this problem, adaptive variance is introduced to judge the convergence of particles using the variance of adaptation, which is calculated as shown in Eq. (18).

$$D = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{f_i - f_{avg}}{f}\right)^2. \tag{18}$$

In Eq. (18), n denotes the population size and f denotes the normalization factor whose value follows the current particle fitness value as shown in Eq. (19).

$$f = \max\left\{1, \max\left\{\left|f_i - f_{avg}\right|\right\}\right\}. \tag{19}$$

When a locally optimal solution occurs in the algorithm, the population extremum G_i varies according to the probability P_z , which is calculated as shown in Eq. (20).

$$P_z = \begin{cases} q, D < \zeta \\ 0, D \geq \zeta \end{cases}. \tag{20}$$

In Eq. (20), ζ denotes a maximum value much smaller than the fitness variance D and q denotes a random number within the interval [0,0.4]. When a particle mutates, it becomes a random number l, which obeys the normal distribution $N(0,1)$, then the random number l is compared with P_z . If $l > P_z$, then the mutation operation is performed, otherwise no mutation is performed, the mutation is shown in Eq. (21).

$$G_i = G_i * (1 + r). \tag{21}$$

After the particle mutation, the probability can leave the local optimal solution and evolve toward the global optimal solution, and the specific flow of the algorithm model is shown in Fig. 3.

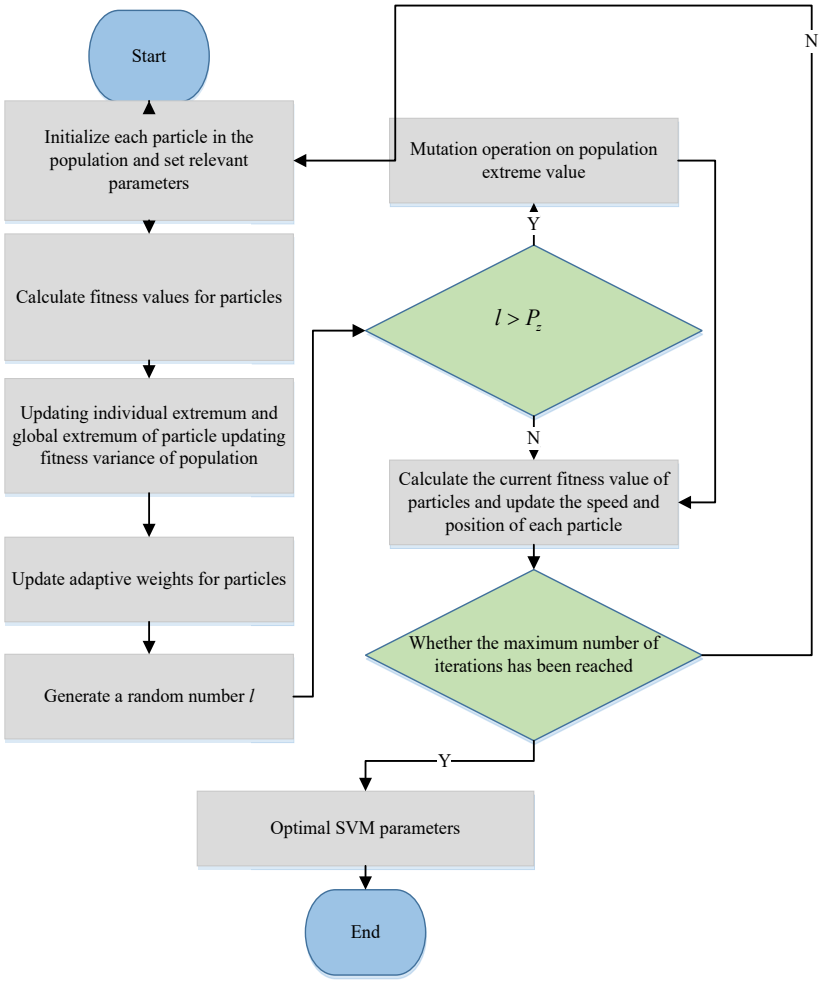

Fig. 3. Flow chart of WPSO-SVM algorithm

As shown in Fig. 3, firstly, the parameters are initialized and the initial positions are randomly generated within a certain range, then the two vectors of the initial positions of the particles are $x(i,1)$, $x(i,2)$, which correspond to the two parameters of the SVM penalty factor C and the kernel parameter σ. Then, the fitness value of each particle is calculated, and the overall fitness variance is calculated based on the fitness values to the individual extremes P_i and the population extremes G_i of the particles. Then, the adaptive weights of each particle are updated based on these data. Then mutation is performed for some particles that satisfy the conditions. Finally, we judge whether the number of iterations is reached, and continue iterating if it is not satisfied, and if it is satisfied, we output the population optimal position, i.e., the optimal SVM parameters.

3.3 Intelligent Home Environment Monitoring System Data Analysis and Processing

The Internet of Things (IoT) denotes an Internet connected between objects, i.e., objects are connected to the Internet through sensing devices to generate connections for intelligent management and identification [19]. This study addresses the problem of automatic feature classification method in the IoT big data environment and selects the most relevant contemporary smart home environment monitoring system as the research object, whose structure is shown in Fig. 4.

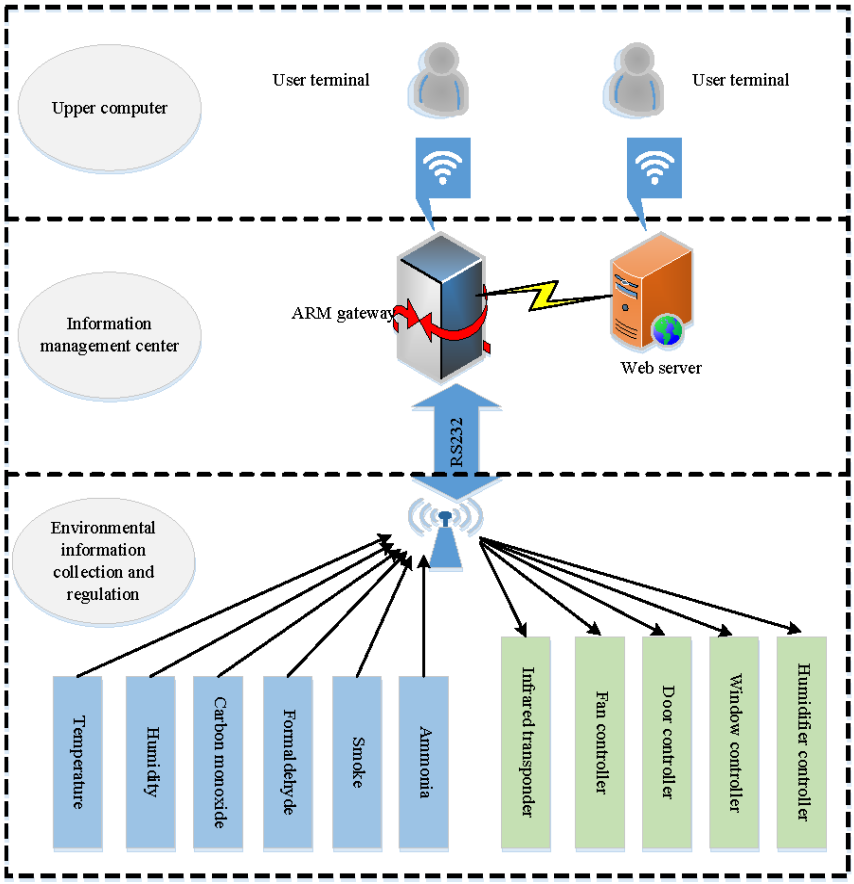

Fig. 4. Intelligent home environment monitoring system

The algorithm model proposed in this study is applied to this smart home environment detection system, which is responsible for feature extraction of the environmental information data collected by the environment collector. Six major environmental factors affecting human health and comfort were selected as the objects for data set construction, namely, temperature, relative humidity, CO concentration, soot concentration, formaldehyde concentration, and ammonia concentration. As shown in Table 1, 10 sets of raw data from a home sensor were extracted as data processing examples.

Table 1. Some raw data collected by the sensor

Serial number	Temperature (℃)	Relative humidity (%)	Carbon monoxide concentration (mg/ m^3)	Dust concentration (mg/ m^3)	Formaldehyde concentration (mg/ m^3)	Ammonia concentration (mg/ m^3)
1	22.3	55.6	10.0	0.17	0.19	0.11
2	23.1	38.7	8.8	0.14	0.09	0.21
3	16.5	58.1	9.4	0.16	0.08	0.22
4	21.2	30.8	8.9	0.22	0.20	0.15
5	22.4	60.3	9.1	0.16	0.12	0.16
6	26.5	72.1	11.2	0.17	0.16	0.31
7	16.7	49.2	7.7	0.19	0.07	0.22
8	22.8	59.2	9.2	0.18	0.16	0.39
9	21.8	61.5	9.3	0.19	0.25	0.19
10	23.9	48.7	7.5	0.22	0.15	0.13

As shown in Table 1, the measurement units and the range of values of different environmental factors data are different, and each environmental factor has a different degree of influence on the environment, which is extremely detrimental to the convergence of the algorithm operation. On the other hand, in the big data environment, the amount of user data and its huge, as the computing model keeps running on, will waste a lot

of time, and at the same time, it is also easy to generate a lot of errors. In order to be able to process the data information better, the data needs to be normalized. The research uses linear function transformation to normalize the original environmental data information collected by the collector. Taking the 10 groups of data in Table 1 as an example, the normalized data is shown in Table 2.

Table 2. Partially normalized environmental data

Serial number	Temperature (℃)	Relative humidity (%)	Carbon monoxide concentration (mg/ m^3)	Dust concentration (mg/ m^3)	Formaldehyde concentration (mg/ m^3)	Ammonia concentration (mg/ m^3)
1	0.51	0.69	0.66	0.02	0.71	0.03
2	0.57	0.22	0.34	0.00	0.24	0.38
3	0.08	0.76	0.50	0.02	0.19	0.41
4	0.43	0.00	0.37	0.06	0.76	0.17
5	0.52	0.82	0.42	0.02	0.38	0.21
6	0.83	1.15	0.97	0.02	0.57	0.72
7	0.10	0.51	0.05	0.04	0.14	0.41
8	0.55	0.79	0.45	0.03	0.57	1.00
9	0.48	0.86	0.47	0.04	1.00	0.31
10	0.63	0.50	0.00	0.06	0.52	0.10

As shown in Table 2, normalizing the data to the [0,1] interval, the processed data is convenient for operation, which is beneficial to improve the speed and performance of the algorithm model and can effectively degrade the error rate.

4 Experiment and Analysis of WPSO-SVM Algorithm Model

The experiments were conducted by using MATLAB, and the Wine dataset was selected from the UCI dataset to compare the three algorithmic models of SVM, PSO-SVM, and WPSO-SVM. The Wine dataset has 178 samples with 13 attributes in a single sample. They are divided into a total of 3 categories, whose category 1 samples have 59, category 2 samples have 71, and category 3 samples have 48. The category 3 samples have 48 samples. The same parameters are set for the algorithmic models, the number of iterations is set to 200, the population size is set to 20, the learning factor is set to $c_1 = 1.5$, $c_2 = 1.7$, the inertia weight is set to $w_{max} = 0.9$, $w_{min} = 0.4$, the penalty factor is set to $0.01 \leq c \leq 100$, and the kernel parameters are set to $0.01 \leq \sigma \leq 10$. The convergence of the fitness of the two algorithmic models, PSO-SVM and WPSO-SVM, for the particles on the Wine dataset is shown in Fig. 5.

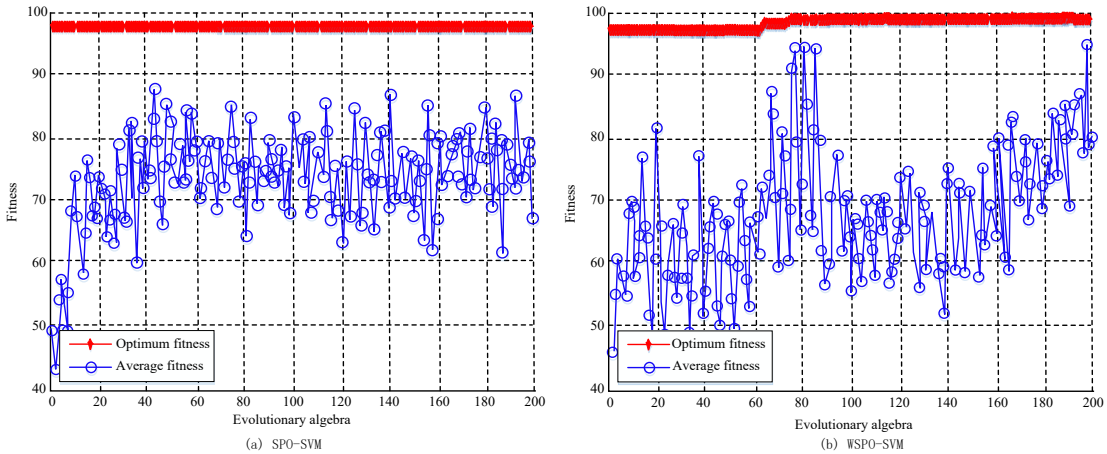

Fig. 5. Fitness convergence of particles

The best and average fitness of the WSPO-SVM algorithm model is higher than that of the SPO-SVM algorithm model. We continue to compare the misclassification sample size and total misclassification sample size of the algorithm models for each category in the Wine dataset, and cross-validate them on the training set to compare their average classification accuracy (CV) and classification precision. The commonly used classification algorithms Incremental Extreme Learning Machine (I-ELM) and BP neuron algorithm were also added to the experiments, and the experimental results are shown in Table 3.

Table 3. Partially normalized environmental data

Algorithmic model	Category 1	Category 2	Category 3	Total misclassification	CV (%)	Classification accuracy (%)
SVM	3	4	5	12	97.14	82.86
PSO-SVM	2	1	6	9	98.08	85.71
WPSO-SVM	2	0	4	6	100	91.43
I-ELM	4	5	4	13	96.36	79.68
BP	4	6	6	16	88.64	84.59

As can be seen from Table 3, the BP neuron algorithm has the largest number of misclassified samples and the WPSO-SVM algorithm has the smallest number of misclassified samples in the Wine data set experiment; the BP neuron algorithm has the lowest accuracy in the average classification of cross-validation and the WPSO-SVM algorithm has the highest accuracy of 100% in the average classification of cross-validation. i-ELM has the set has the lowest classification accuracy and the WPSO-SVM algorithm has the highest classification accuracy. From the experimental results, it can be seen that the overall accuracy and precision of SVM-like algorithms are higher than other algorithms. Among the similar algorithms, the optimized PSO-SVM algorithm and WPSO-SVM algorithm also have higher accuracy and precision than the SVM algorithm, indicating that the optimization is effective and meets expectations. The WPSO-SVM algorithm model optimized again on the PSO-SVM optimized algorithm model performs more outstandingly and outperforms other algorithms in all aspects, which proves that the optimization is effective and meets expectations. wine dataset is relatively simple compared to the dataset to be processed in the IoT big data environment, and in order to further study the performance of the algorithm, the data from the smart home environment monitoring system is extracted as Home Environment Dataset (Hed) for experimentation. Thirty sets of home environment data were randomly selected and divided into training sample set and test sample set according to a certain ratio. According to the different types of data, the data are divided into three categories, category 1 has 6 sets of samples, set 5 training groups and 1 test group; category 2 has 7 sets of samples, set 5 training groups and 2 test groups; category 3 has 17 sets of samples, set 11 training groups and 6 test groups. In total, there are 30 groups of samples, of which 21 are training groups and 9 are testing groups. The data of the selected 21 training groups are input into the WPSO-SVM model for training, and the number of iterations is changed to 100, and the rest parameters are kept unchanged, and the obtained particle fitness convergence is shown in Fig. 6.

As can be seen from Fig. 6, the actual adaptation gradually tends to be optimal as the number of iterations increases, and the optimal adaptation reaches the optimum when it evolves to about 23 generations, i.e., the optimal parameters of SVM are found. In order to more accurately compare the processing results of each algorithm under the Hed data set, 300 sets of sample data were randomly selected in the smart home environment system, and a total of 10 sets of data were used to compare the classification accuracy of each of the five algorithms according to the way of 30 sets as one data set, and the obtained experimental results are shown in Fig. 7.

As can be seen from Fig. 7, the WSPO-SVM algorithm model still performs the best under the smart home environment monitoring data, reaching 100% accuracy in 6 out of 10 sets of test data. The BP neuron algorithm performs the worst, with an average accuracy of 80% under 10 sets of test data. The 5 algorithm models, whose average accuracy ranking is WPSO-SVM, PSO-SVM SVM, I-ELM, and BP, were 97.67%, 94.44%, 88.89%, 83.33%, and 80%, respectively. The experimental results show that SVM is more accurate than ELM and BP models in smart home environment monitoring. The proposed particle swarm optimization algorithm has about 5.5% capacity improvement on the SVM model, and the proposed adaptive optimization has about 7.8% capacity improvement on the SVM model. The experimental results show that the proposed WPSO-SVM algorithm model can be well applied to the smart home environment monitoring system, and is superior to other algorithms in accuracy and stability.

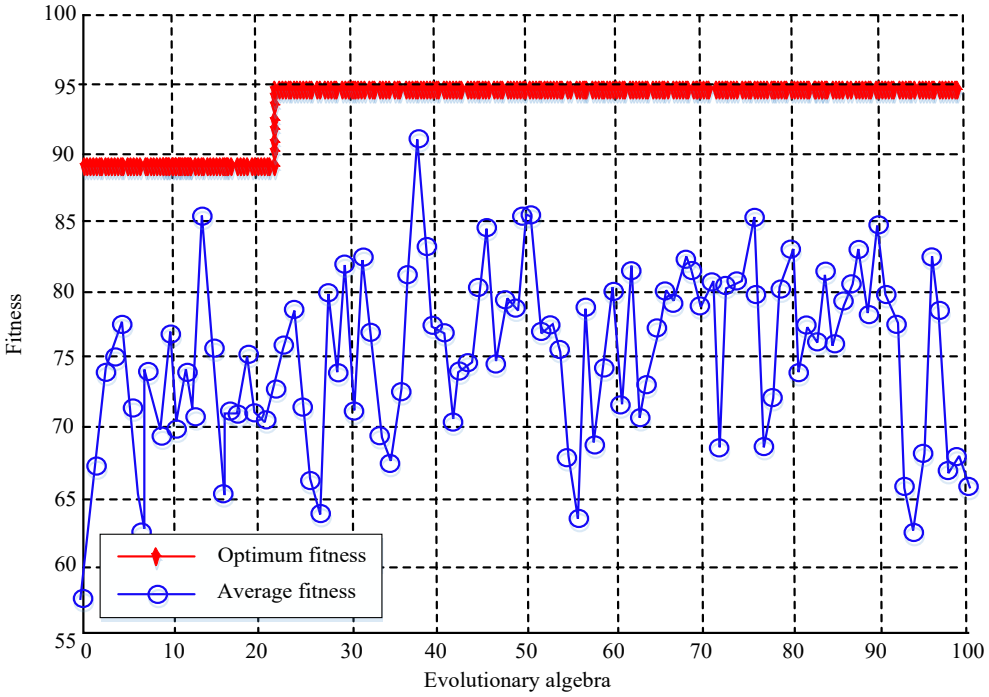

Fig. 6. Particle convergence of wspo-svm model

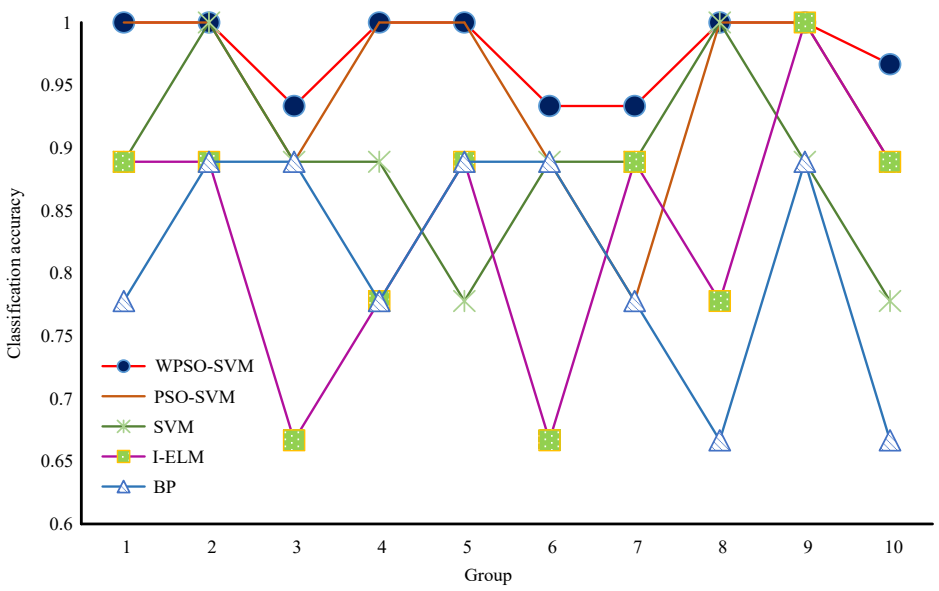

Fig. 7. Comparison chart of algorithm accuracy

5 Conclusion

This study addresses the problem of automatic classification method for IoT big data features using SVM

algorithm model, optimizes the SVM algorithm, proposes the SPO-SVM algorithm model, and proposes the WSPO-SVM algorithm model by adaptive optimization on the SPO-SVM model for the problem that the SPO algorithm is easy to fall into local optimal solutions. The results show that the WSPO-SVM model has less total misclassification and single class misclassification compared with other algorithms under the Wine dataset. In cross-validation, its classification accuracy performance is better than other algorithms in both cases, indicating that the WSPO-SVM algorithm outperforms other algorithms in terms of accuracy and achieves the expected goal. Under 10 sets of smart home environment monitoring data sets, the WSPO-SVM algorithm model still performs the best, with 100% accuracy in 6 out of 10 test data sets, with an average accuracy of 97.67%, which is about 9% higher than the common SVM algorithm model and about 15% higher than other feature classification algorithms. The experimental results prove that the WSPO-SVM algorithm can complete the automatic feature classification work in the large data environment of IoT, which meets the expectation. Of course, there are some shortcomings in this study, such as the experimental data set does not have enough samples and is not complex enough, and the application environment of IoT is not rich enough to be selected. It is expected that in future research, more data sets can be used for experiments to improve the accuracy of the experiments, and more IoT environments can be adapted to improve the practicality of the research.

References

[1] K. Ghaffari, M. Lagzian, M. Kazemi, G. Malekzadeh, A socio-technical analysis of internet of things development: an interplay of technologies, tasks, structures and actors, Foresight 21(6)(2019) 640-653.

[2] C. Yao, Y. Li, Research on knowledge innovation framework based on internet of things and big data, Journal of Physics: Conference Series 1992(3)(2021) 032046.

[3] R. Bai, J.-Y. Lv, C. Shang, Adaptive cognitive management and knowledge discovery framework based on internet of things big data, Journal of Physics: Conference Series 1802(4)(2021) 042080.

[4] V. Zhmud, A. Liapidevskiy, V. Avrmachuk, V. Sayapin, O. Stukach, H. Roth, Analysis of barriers to the development of industrial internet of things technology and ways to overcome them, IOP Conference Series: Materials Science and Engineering 1019(1)(2021) 012079.

[5] X. Lv, M. Li, Application and research of the intelligent management system based on internet of things technology in the era of big data, Mobile Information Systems 2021(2021) 6515792.

[6] N.A. Al-Thanoon, Z.Y. Algamal, O.S. Qasim, Feature selection based on a crow search algorithm for big data classification, Chemometrics and Intelligent Laboratory Systems 212(2021) 104288.

[7] Y. Wang, L. Yang, A robust loss function for classification with imbalanced datasets, Neurocomputing 331(2019) 40-49.

[8] E. Sevinc, A novel evolutionary algorithm for data classification problem with extreme learning machines, IEEE Access 7(2019) 122419- 122427.

[9] T. Tang, S. Chen, M. Zhao, W. Huang, J. Luo, Very large-scale data classification based on K-means clustering and multi-kernel SVM, Soft Computing - A Fusion of Foundations, Methodologies and Applications 23(11)(2019) 3793-3801.

[10] B. Gaye, D. Zhang, A. Wulamu, Improvement of support vector machine algorithm in big data background, Mathematical Problems in Engineering 2021(2021) 5594899.

[11] J. Balakumar, S.V. Mohan, Artificial bee colony algorithm for feature selection and improved support vector machine for text classification, Information Discovery and Delivery 47(3)(2019) 154-170.

[12] J. Wei, H. Huang, P.-D. Kang, PSO-DEC-IFSVM classification algorithm for unbalanced data, Shu Ju Cai Ji Yu Chu Li/ Journal of Data Acquisition and Processing 34(4)(2019) 723-735.

[13] X. Yin, Y.-D. Hou, J. Yin, C. Li, A novel SVM parameter tuning method based on advanced whale optimization algorithm, Journal of Physics: Conference Series 1237(2)(2019) 022140.

[14] D. Saputra, W.-S. Dharmawan, M. Wahyudi, W. Irmayani, J. Sidauruk, Martias, Performance comparison and optimized algorithm classification, Journal of Physics: Conference Series 1641(1)(2020) 012087.

[15] D.J. Kalita, V.-P. Singh, V. Kumar, A dynamic framework for tuning SVM hyper parameters based on moth-flame optimization and knowledge-based-search, Expert Systems with Applications 168(2020) 114139.

[16] I.N.-A. Firdausanti, On the comparison of crazy particle swarm optimization and advanced binary ant colony optimization for feature selection on high-dimensional data, Procedia Computer Science 161(2019) 638-646.

[17] J. Marietta, B.-C. Mohan, A review on routing in internet of things, Wireless Personal Communications 111(1)(2020) 209-233.

[18] O.-F. Aje, A.-A. Josephat, The particle swarm optimization (PSO) algorithm application - A review, Global Journal of Engineering and Technology Advances 3(3)(2020) 1-6.

[19] X. Li, N. Zhao, R. Jin, S. Liu, X. Sun, X. Wen, D. Wu, Y. Zhou, J, Guo, S. Chen, Z. Xu, M. Ma, T. Wang, Y. Qu, X. Wang, F. Wu, Y. Zhou, Internet of things to network smart devices for ecosystem monitoring, Science Bulletin 64(17) (2019) 1234-1245.