IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON HUMAN-CENTERED SMART SYSTEMS AND TECHNOLOGIES

Received January 24, 2018, accepted February 22, 2018, date of publication March 5, 2018, date of current version March 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2810198

An Improved Intrusion Detection Algorithm Based on GA and SVM

PEIYING TAO, ZHE SUN, AND ZHIXIN SUN

Key Laboratory of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210023, China

Corresponding author: Zhixin Sun (sunzx@njupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61373135, Grant 61672299, Grant 61702281, and Grant 61602259, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20150866 and Grant BK20160913.

ABSTRACT In the era of big data, with the increasing number of audit data features, human-centered smart intrusion detection system performance is decreasing in training time and classification accuracy, and many support vector machine (SVM)-based intrusion detection algorithms have been widely used to identify an intrusion quickly and accurately. This paper proposes the FWP-SVM-genetic algorithm (GA) (feature selection, weight, and parameter optimization of support vector machine based on the genetic algorithm) based on the characteristics of the GA and the SVM algorithm. The algorithm first optimizes the crossover probability and mutation probability of GA according to the population evolution algebra and fitness value; then, it subsequently uses a feature selection method based on the genetic algorithm with an innovation in the fitness function that decreases the SVM error rate and increases the true positive rate. Finally, according to the optimal feature subset, the feature weights and parameters of SVM are simultaneously optimized. The simulation results show that the algorithm accelerates the algorithm convergence, increases the true positive rate, decreases the error rate, and shortens the classification time. Compared with other SVM-based intrusion detection algorithms, the detection rate is higher and the false positive and false negative rates are lower.

INDEX TERMS Genetic algorithm, intrusion detection, support vector machine.

I. INTRODUCTION

With the development and popularization of information and network technologies, network information security is becoming more and more important. Compared with traditional network defense technology (such as firewalls), human-centered smart IDSs that can take initiative to intercept and warn of network intrusion has a great practical value. The question of how to improve the effectiveness of smart network intrusion detection has become a focus of network security [1].

Currently, use of smart IDS is viewed as an effective solution for network security and protection against external threats. However, the existing IDS often has a lower detection rate under new attacks and has a high overhead when working with audit data, and thus machine learning methods have been widely applied in intrusion detection. SVM, one of the machine learning technologies, is a new algorithm based on statistical learning theory that has shown higher performance than the traditional learning methods in solving the classification problem of pattern recognition and speech recognition [2]. Compared with other classification algorithms, SVM can better solve the problems of small samples, nonlinearity and high dimensionality. However, with the advent of the era of big data, SVM encounters the problem of long training and testing times, high error rates and low true positive rates, which limit the use of SVM in network intrusion detection. Therefore, SVM feature selection, feature weighting and SVM parameter setting are critical to improved detection performance. GA shows excellent global optimization ability via population search strategies and information exchange between individuals. Different from the traditional multi-point search algorithm, GA can easily avoid local optima. In this paper, GA and SVM are used to select the optimal feature subset and optimize the SVM parameters and feature weights to improve the performance of the network intrusion detection system.

The remainder of the paper is organized as follows. Section II describes related work, Section III introduces

the genetic algorithms (including selection operators, and optimized crossover and mutation probability), Section IV presents an improved intrusion detection method based on GA and SVM (including selection of the optimal feature subset and optimization of the SVM parameters and feature weighting), Section V verifies the effectiveness of the FWP-SVM-GA algorithm by comparing the experimental results with other methods of intrusion detection, and Section VI presents conclusions.

II. RELATED WORK

In the era of big data, intrusion detection has become the most important topic in security infrastructure. To distinguish between attack and normal network access, different machine learning methods are applied in IDS, including fuzzy logic [3], K nearest neighbor (KNN) [4], support vector machine (SVM), artificial neural network (ANN) [5], and artificial immune system (AIM) approaches [6]. SVM showed better performance than traditional classification techniques [7], and several researchers proposed SVM-based IDS [8]–[10]. Although SVM-based IDS can improve IDS performance in terms of detection rate and learning speed compared with traditional algorithms (such as neural networks), room for improvement still exists. As the number of features of the audit data becomes larger, the performance of IDS degrades in terms of training time and classification accuracy. To address these problems, we use GA technology to supply fast and accurate optimization that can enable IDS to find the optimal detection model based on SVM.

In [11], the genetic algorithm (GA) was proposed to improve the intrusion detection system (IDS) based on support vector machine (SVM), and the optimal feature subset was selected for SVM. However, the error rate of SVM was not considered. In [12], an intrusion detection method based on wavelet kernel least square was designed to improve the detection capability of SVM in complex nonlinear systems. However, the training and testing time of the algorithm is relatively long. In [13], the heuristic genetic algorithm was applied to optimize the SVM kernel parameters. The genetic operator is dynamically adjusted via a heuristic strategy, and the classification accuracy of the model is taken as the objective function to realize parameter optimization of the Gaussian kernel-based SVM classification model. However, this approach did not consider the impact of feature weighting on SVM detection accuracy. In [14], the coarse-grained parallel genetic algorithm (CGPGA) was presented to simultaneously optimize the feature subsets and parameters of SVM. A new fitness function was proposed that includes the classification accuracy, the number of features and the number of support vectors, but it required a long time to train the SVM. In [15], GA was selected as one of the most powerful tools to search in a large space with the potential to find the best solution in the search space. However, in the later evolution of the population, a larger crossover and mutation probability might result in the loss of good genes and delayed convergence of the algorithm.

In summary, although many SVM-based network intrusion detection methods have been proposed in recent years, the above algorithms still suffer from certain shortcomings:

- Due to redundant features, the raw dataset confuses the classifier, leading to inaccurate detection. Traditional feature selection (such as PCA) ignores a number of sensitive features, resulting in a classifier without optimal sensitivity.
- If GA is used to optimize the SVM-based intrusion detection system, the training time is longer, and the error rate is higher when selecting the optimal feature subset. After selecting the optimal feature subset, the importance of the features is not sorted.

For these reasons, we propose a combination of the genetic algorithm (GA) with support vector machine (SVM). First, we optimize the crossover probability and mutation probability of GA, generate the population to speed up the search in the early evolution of the population and accelerate the convergence of the algorithm in the later evolution of the population. In the stage of optimal feature set selection, a new fitness function is proposed to decrease the error rate while increasing the true positive rate. Finally, the feature weights and parameters of SVM are optimized simultaneously, and the robustness of SVM is improved.

III. GENETIC ALGORITHMS

Genetic operators are the key to optimization, and specifically, the crossover and mutation operators are used to maintain population diversity and avoid local optima. Currently, the crossover probability and mutation probability are constants during the period of population evolution and delay the convergence of the algorithm in the later evolution of the population, leading to the long training time of SVM. Therefore, the method proposed in this paper changes the crossover probability and mutation probability of GA according to the evolutionary algebra and fitness value, which generates the population to speed up the search in the early evolution of the population and accelerates the convergence of the algorithm in the later evolution of the population.

A. SELECTION OPERATORS

The selection in GA is designed to seek better individuals and maintain the diversity of the population. The offspring population chooses the individual using the fitness value, which gives the higher quality individual a greater chance to be chosen. The common selection operators are roulette wheel selection, elitist selection, and tournament selection.

Roulette wheel selection: Selection of a chromosome in the population is proportional to its fitness value. The population is assigned a circular ‘‘roulette wheel" slice, which is proportional to the individual’s fitness value, and the wheel rotates N times (N is the number of individuals in the population). In each rotation, the chromosome under the wheel mark is selected in the next generation.

Elitist selection: The individual with the highest fitness value in the population does not participate in

crossover or mutation and is used to replace the individual with the lowest fitness value after crossing and mutation. Elitist selection avoids loss of the optimal individual by the crossover or mutation operator.

Tournament selection: The selection process runs a number of ''tournaments" between two individuals randomly selected from the population, and the better individual with a greater fitness value is selected for the next generation. In this paper, the optimal 60 percent of chromosomes was selected using the tournament selection method.

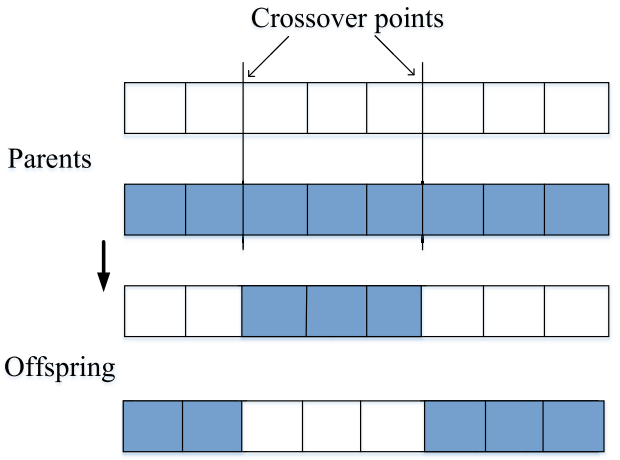

FIGURE 1. Crossover operators and their effects in generation of the offspring.

B. OPTIMIZED CROSSOVER PROBABILITY

The partial structure of the parent chromosomes is replaced and recombined to form a new individual, and this operation is referred to as the crossover operation, as shown in Fig. 1. With the increase in the population evolutionary algebra, the population approaches the optimal solution set, and thus we stress the following points.

In the early evolution of the population, it is necessary to increase the number of individual crossovers for rapid search over the whole definition space. At the later stage of population evolution, the population is concentrated in the vicinity of the optimal solution, and the number of individual crossovers must be reduced to prevent the loss of individual good genes to speed up GA convergence.

When the average fitness value of individuals is low, an increase in the individual crossover probability increases the possibility of generating excellent individuals. When the average fitness value of population approaches the optimal solution, the individual crossover probability should be reduced.

In summary, the adaptive crossover probability in the genetic operation is given as follows:

$$P_c = \frac{P_{c0} + P_{c1}}{2} = \frac{\left(\left(\frac{N-n}{N}P_{cmax} + \frac{n}{N}P_{cmin}\right) + \frac{P_{cmax} f_{min}}{f_{max}}\right)}{2} \quad (1)$$

With the increase of population evolutionary algebra, the value of P_{c0} decreases, and P_{c1} decreases when the average fitness value of the population tends toward the ideal value. In this case, P_{cmax} is the maximum crossover probability, P_{cmin} is the minimum crossover probability, f_{min} is the minimum fitness value of the cut-off for the current population, f_{max} is the maximum fitness value of the cut-off for the current population, n is the current evolutionary algebra, and N is the evolutionary algebra of the entire population.

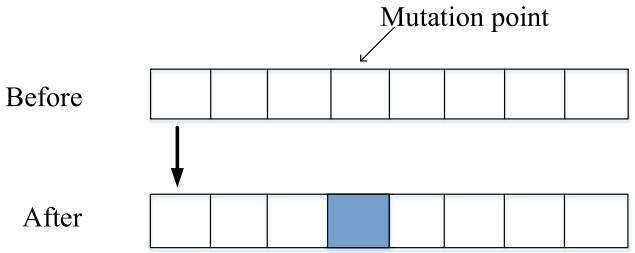

FIGURE 2. Mutation operators and their effects in generation of the offspring.

C. OPTIMIZED MUTATION PROBABILITY

Mutations can alter individual chromosomal genes in the parent population, resulting in a large number of new individual offspring, as shown in Fig. 2. With the increase in evolutionary algebra, the population grows approaches the optimal solution set. Choosing a larger probability of mutation produces many new individuals. These new individuals are distributed throughout the search space, and the proportion of individuals with good fitness in the population declines. Therefore, at the later stage of the population evolution, a larger probability of mutation affects the proportion of the dominant individuals and delays the convergence of the algorithm. Therefore, the mutation probability is given as follows:

$$P_m = \frac{N-n}{N}P_{mmax} + \frac{n}{N}P_{mmin} \quad (2)$$

where P_{mmax} is the maximum mutation probability, and P_{mmin} is the minimum mutation probability.

IV. IMPROVED INTRUSION DETECTION METHOD BASED ON GA AND SVM(FWP-SVM-GA)

Fig. 3 shows the architecture of the improved intrusion detection method based on GA and SVM for feature selection, feature weighting and SVM parameter optimization, as proposed in this paper. The input is the network traffic data set, and the final output is attack detection and alarm. The system consists of four main components:

- Feature selection based on GA and SVM: The network traffic data are entered, feature chromosomes are created, the chromosomes according to the fitness function proposed in this paper are evaluated, the chromosomes with the maximum fitness function value as the optimal

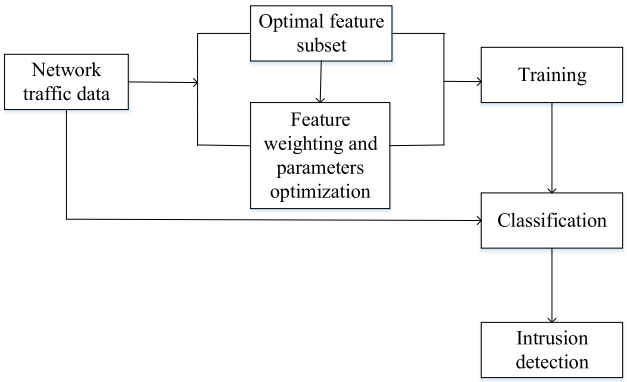

FIGURE 3. System architecture of the proposed FWP-GA-SVM IDS.

chromosome are selected, and the optimal feature subset is decoded.

- Feature weighting and parameter optimization based on GA and SVM: The weights of the feature and SVM parameter chromosomes are created according to the optimal feature subset. By evaluating the chromosome with the highest classification accuracy and selecting it as the optimal chromosome, the optimal SVM parameters and the feature weights are decoded.
- Training: The original data are randomly divided into k sub-portions of the same size to retain the first, second, ... k sub-portions, and the remaining k-1 sub-portions are used as training data to training the support vector machine.
- Classification: The reserved first, second, ... k sub-portions are classified as testing data and the combined k prediction results. The advantage of this technique is that all testing sets are independent and can improve the reliability of the results. In our experiments, we chose $k = 10$ and combine k results to estimate the performance of the SVM classifier.

A. FEATURE SELECTION BASED ON GENETIC ALGORITHM (GA) AND SUPPORT VECTOR MACHINE (SVM) (FWP-SVM-GA-1)

1) FLOW CHART (FIG. 4)

2) STEPS

- a. The feature chromosomes and original data are generated.
- b. Using the decoding parameters and training datasets to create the SVM intrusion detection model, the testing datasets are used to evaluate the classifier (K-fold cross validation), and each chromosome is evaluated using the fitness function when the prediction results are obtained.
- c. To evaluate the GA termination conditions, we determine whether the maximum number of 100 iterations is reached or the current generation of the maximum fitness value minus the previous generation of the maximum fitness value is less than 0.001. If so, we jump to step e, otherwise, we move to the next step.

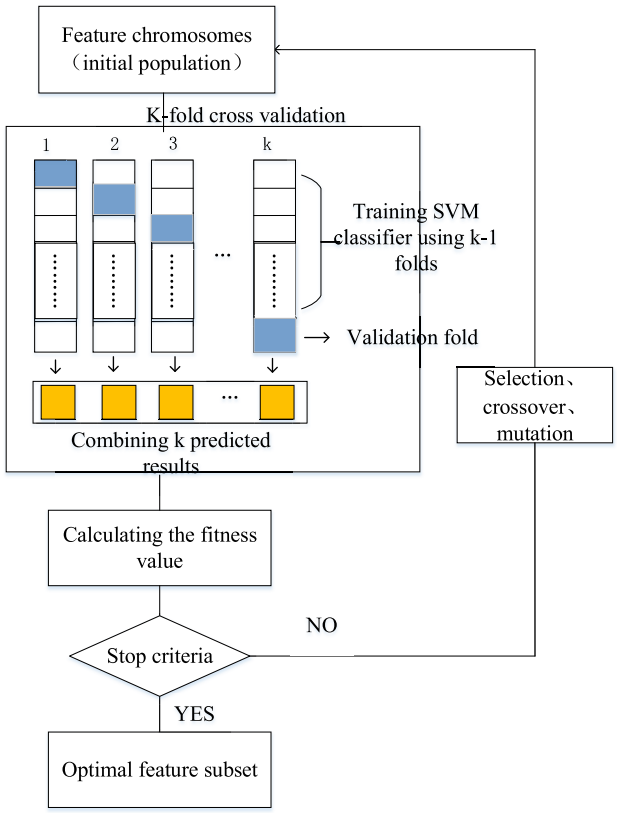

FIGURE 4. Feature selection based on genetic algorithm (GA) and support vector machine (SVM).

- d. Genetic operations (selection, crossover and mutation) are performed on the parent population, and the offspring population is generated. The operation returns to step b.
- e. The optimal feature subset is obtained by decoding the chromosome that has the maximum fitness function value.

3) CHROMOSOME DESIGN

The chromosome design is binary coded, where 1 indicates that the feature index is selected, and 0 indicates that the feature index is not selected. For example: feature chromosome 1: 110011110010 ... 10001.

4) PROPOSED FITNESS FUNCTION

The fitness function is a basic component in GA that can evaluate whether an individual is suitable for survival. In this paper, a new fitness function is proposed for GA to decrease the error rate and increase the true positive rate while choosing the optimal feature subset.

The new fitness function evaluates each feature subset using three parameters, i.e., the true positive rate (TPR), error rate (Error) and number of selected features (NumF (S)). The calculation formula is written as follows:

$$\text{Fitness (S)} = W_a^{'} TPR + W_b^{'} Error + W_c^{'} NumF\ (S) \quad (3)$$

The true positive rate (TPR) refers to the proportion of samples that the classifier correctly predicts in all samples for which the actual category is positive, and the calculation formula is given as follows:

$$\text{True positive rate (TPR)} = \text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

The error rate (Error) refers to the proportion of the samples that the classifier incorrectly predicts in all samples, and the calculation formula is presented as follows:

$$\text{Error rate (Error)} = \frac{FP + FN}{TP + FN + TN + FP} \tag{5}$$

Where Wa is the weight value of TPR, Wb is the weight value of Error, and Wc is the weight value of the selected feature number. In general, Wa and Wb can be set from 75% to 100% according to the needs of the user. In this paper, Wa is set to 40 %, Wb to 50 %, and Wc to 10 %, to obtain the highest TPR, the lowest Error, and the smallest feature subset. The confusion matrix is shown in Table 1, which contains the actual and predicted classification information produced by the classification system.

TABLE 1. Confusion matrix.

		Predicted	Predicted
		1	-1
Actual	1	True Positive (TP)	False Negative (FN)
Actual	-1	False Positive (FP)	True Negative (TN)

B. PARAMETERS OPTIMIZATION AND DATA FEATURE WEIGHTING BASED ON GENETIC ALGORITHM (GA) AND SUPPORT VECTOR MACHINE (SVM) (FWP-SVM-GA-2)

After selecting the optimal feature subset, SVM also faces two problems: how to sort the importance of the feature and how to choose the optimal SVM parameters. These two issues must be resolved at the same time because the weighting feature influences the kernel parameter and vice versa.

1) FLOW CHART (FIG. 5)

2) STEPS

- a. Generation of chromosomes and original data.
- b. Decoding and data transformation: The chromosomes are converted into SVM parameters C, γ(Eq. (7), (8)) and feature weights. In the training and testing data sets, we multiply the feature value of an instances by the corresponding weights using Eq. (6).Where N_{xy} and M_{xy} are the values of the y^{th} field of the x^{th} instance before and after the transformation, and W_y is the weight of the y^{th} field.
- c. The C, γ and transformed training data sets are used to construct the SVM model. The transformed testing data sets are used to evaluate the performance of the classifier (K-fold cross validation). When the predicted results are obtained, each chromosome is evaluated using the fitness function.

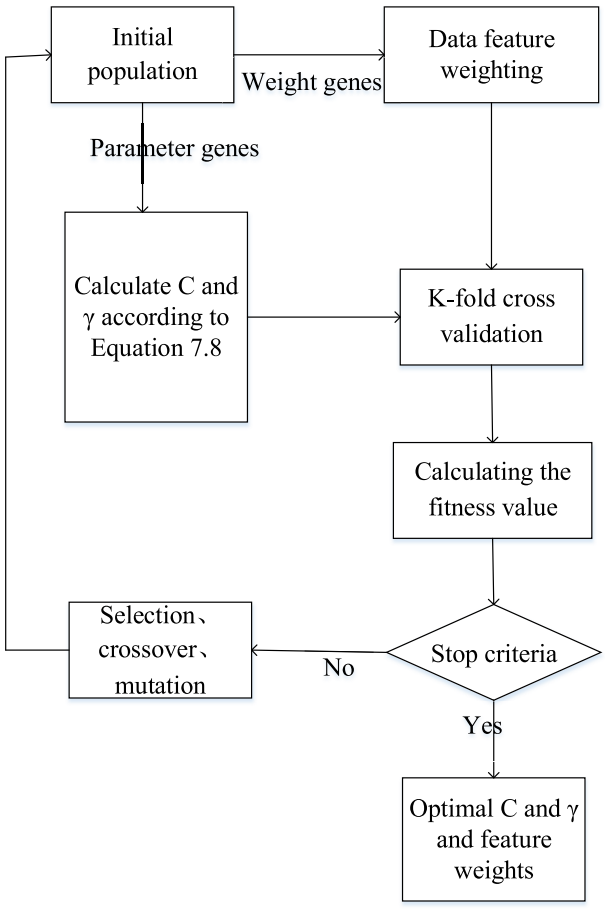

FIGURE 5. Parameter optimization and data feature weighting based on genetic algorithm (GA) and support vector machine (SVM).

- d. We determine whether to meet GA termination conditions. If yes, we jump to step f, otherwise, the process continues to the next step.
- e. Genetic operations (selection, crossover and mutation) are performed on the parent population, the offspring population is generated, and the process jumps to step b.
- f. The chromosome with the maximum fitness value is decoded to obtain the feature weights and the optimal SVM parameters.

$$M_{xy} = N_{xy} * W_y \tag{6}$$

3) CHROMOSOME DESIGN

The chromosome design is real-number coded. The RBF kernel function of the SVM is used to convert a completely non-separable problem into a separable or approximate separable state. The RBF kernel parameter γ implies the distribution of the data to the new feature space. The parameter C represents the degree of penalty for the classification error in the linear non-separable case. Because the weighting feature and the support vector machine parameters interact with each other, the chromosome must include both the parameters and the feature weights, as shown in Fig. 6.

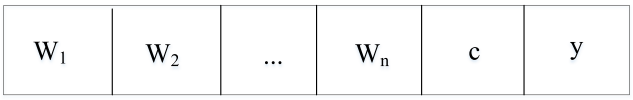

FIGURE 6. Structure of chromosome of C, γ and feature weights.

All genes in the chromosome have values in the range [0,1]. The two genes c and y represent the gene values of C and γ, and W1 through Wn represent the gene values of the feature weights (the weight of the unselected feature W = 0). Thus, the SVM parameters C and γ map c and y to [C1, C2] and [γ1, γ2] to obtain the formulas as follows:

$$C = C_1 + c * (C_2 - C_1) \tag{7}$$

$$\gamma = \gamma_1 + y^* (\gamma_2 - \gamma_1) \tag{8}$$

4) FITNESS FUNCTION

In the decoding process, the i^{th} feature of the training and testing datasets is multiplied by the corresponding weights W_i ($i = 1, \ldots, n$), and the SVM with the RBF kernel function is built based on C, γ and the transformed training datasets. The classification accuracy of the testing datasets is used to assess the quality of the chromosome. The fitness function is expressed in terms of accuracy, and the calculation formula is given as follows:

$$\text{Fittness} = \text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{9}$$

V. SIMULATION

A. SIMULATION ENVIRONMENT

In this experiment, the simulation model is built on Matlab 2014a software. The datasets are sourced from KDD Cup 99, and the parameter settings are shown in Table 2.

TABLE 2. Parameters of GA and SVM.

Parameter	Value
Number of generations-N	100
Size of population	500
Maximum crossover probability-Pcmax	0.9
Minimum crossover probability-Pcmin	0.4
Crossover type	Two point
Maximum mutation probability-Pmmax	0.1
Minimum mutation probability-Pmmin	0.0001
Search range of parameter C	[0.001,1000]
Search range of kernel function parameter γ	[0.00001,64]
Mutation type	Simple mutation

B. EXPERIMENTAL RESULTS AND DISCUSSION

1) INFLUENCE OF OPTIMIZED GENETIC OPERATORS ON SVM CONVERGENCE

The optimized crossover and mutation probabilities vary with the number of iterations, and the crossover and mutation probabilities take on constant values when not optimized. Fig. 7 contrasts the SVM convergence algebra between the two values. The x-axis represents the number of iterations, and the y-axis denotes the accuracy of SVM (Accuracy).

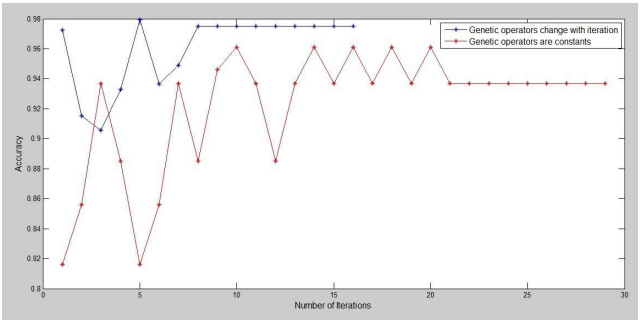

FIGURE 7. Comparison of convergence speed of SVM.

It can be observed from Fig. 7 that the SVM of the optimized genetic operators converges faster than those that are not optimized.

TABLE 3. Features of the KDD Cup 99 datasets.

No.	Feature name	No.	Feature name
1	duration	22	is_guest_login
2	protocol_type	23	Count
3	service	24	serror_rate
4	src_byte	25	rerror_rate
5	dst_byte	26	same_srv_rate
6	flag	27	diff_srv_rate
7	land	28	srv_count
8	wrong_fragment	29	srv_serror_rate
9	urgent	30	srv_rerror_rate
10	hot	31	srv_diff_host_rate
11	num_failed_logins	32	dst_host_count
12	logged_in	33	dst_host_srv_count
13	num_compromised	34	dst_host_same_srv_count
14	root_shell	35	dst_host_diff_srv_count
15	su_attempted	36	dst_host_same_src_port_rate
16	num_root	37	dst_host_srv_diff_host_rate
17	num_file_creations	38	dst_host_serror_rate
18	num_shells	39	dst_host_srv_serror_rate
19	num_access_shells	40	dst_host_rerror_rate
20	num_outbound_cmds	41	dst_host_srv_rerror_rate
21	is_hot_login		

2) IMPROVED ALGORITHM FOR SELECTION OF THE OPTIMAL FEATURE SUBSET (FWP-SVM-GA-1)

The support vector machine (SVM) classifies datasets by selecting important features. In this paper, based on the genetic algorithm and support vector machine, 19 important features are selected. From the experiments using these features, as shown in Table 5, it can be observed that feature selection reduces the classification time and increases the classification accuracy. The improvement in classification accuracy is due to the elimination of confusion caused by irrelevant attributes by reducing the features. In addition, because the number of rules for decision making is reduced, the classification time is reduced. The KDD Cup 99 datasets contain the features shown in Table 3. The features selected in this paper are shown in Table 4.

3) SIMULATION RESULTS OF IMPROVED FITNESS FUNCTION WHEN SELECTING OPTIMAL FEATURE SUBSET

The choice of the appropriate fitness function plays a highly important role in the genetic algorithm (GA). The fitness

TABLE 4. Feature selection based on GA and SVM.

Selected features
duration,protocol_type,dst_byte,urgent,su_attempted,num_root,num_file_creations, num_access_shells,num_outbound_cmds,is_hot_login,rerror_rate,same_srv_rate, srv_count,srv_serror_rate,srv_rerror_rate,dst_host_diff_srv_count, dst_host_srv_diff_host_rate,dst_host_srv_serror_rate,dst_host_srv_rerror_rate

TABLE 5. Comparison of support vector machine (SVM) performance for selecting the optimal feature subset and not selecting the optimal feature subset.

Training set (10000)		
	Accuracy	Classification time
SVM (41 features)	0.9956	5.1875
FWP-SVM-GA-1 (19 features selected in this paper)	0.9975	5.0781

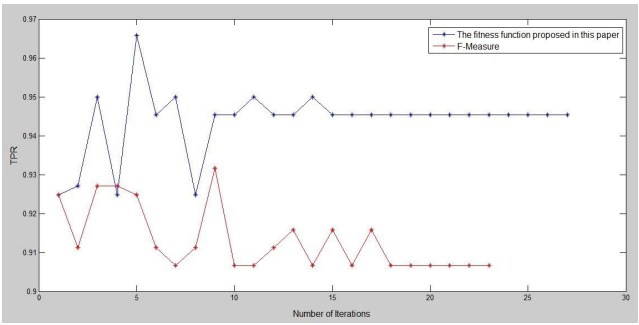

FIGURE 8. Support vector machine (SVM) true positive rate (TPR) contrast between the fitness function proposed in this paper and the fitness function F-measure when selecting the optimal feature subset.

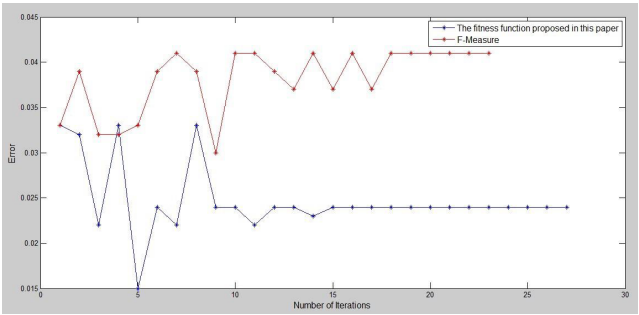

FIGURE 9. Support vector machine (SVM) error rate (Error) contrast between the fitness function proposed in this paper and the fitness function F-measure when selecting the optimal feature subset.

function locates the GA search strategy, which can obtain the best solution in a large search space. The appropriate fitness functions help the GA to explore the search space more efficiently. In contrast, inappropriate fitness functions can cause the GA to fall into local optima solutions easily and thus lose the ability to explore. Compared with the fitness function F-measure, the improved fitness function makes SVM obtain a higher true positive rate (TPR), as shown in Fig. 8, and a lower error rate (Error), as shown in Fig. 9, and the convergence rate is also faster than that of the F-measure. As shown in Table 6, the fitness function proposed in this paper results in a SVM final true positive rate of 94.53 %, an error rate of 2.4 %, convergence algebra of 15, and the fitness function is the

TABLE 6. Support vector machine (SVM) performance contrast between the fitness function proposed in this paper and the fitness function F-measure when selecting the optimal feature subset.

Fitness function	TPR	Error	Number of Iterations
F-measure	0.9066	0.041	18
Fitness function proposed in this paper	0.9453	0.024	15

F-measure (Eq. (10), Precision is obtained by Eq. (11). Recall is obtained by Eq. (4)), making the SVM final true positive rate 90.66 % with an error rate of 4.1 %, and convergence algebra of 18.

$$\mathrm{F-measure} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{10}$$

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{11}$$

4) PERFORMANCE OF THE IMPROVED INTRUSION DETECTION ALGORITHMS BASED ON GA AND SVM (FWP-SVM-GA)

In this paper, we propose two-step optimization of SVM based on GA. The first step is to select the feature subset (FWP-SVM-GA-1) that can result in the SVM with the maximum true positive rate and minimum error rate. The second step is to optimize the selected feature weights and SVM parameters (FWP-SVM-GA-2). Compared with the intrusion detection algorithm, which is based on the ant colony network and support vector machine (SVM) proposed in [16], the detection rate (DR) of each step in the FWP-SVM-GA algorithm is higher than that of CSVAC, and the false positive rate (FPR) and false negative rate (FNR) of each step in the FWP-SVM-GA algorithm are lower than those of CSVAC. Compared with the support vector machine using the heuristic genetic algorithm proposed in [13], the FWP-SVM-GA algorithm has a higher detection rate for each step. Compared with the GF-SVM algorithm proposed in [17], the final false positive rate (FPR) and false negative rate (FNR) of the FWP-SVM-GA algorithm are lower than those of GF-SVM, as shown in Table 7.

TABLE 7. Comparison of performance between FWP-SVM-GA and other algorithms.

Algorithm	DR	FNR	FPR
CSVAC [16]	94.86	1.00	6.01
HGA-SVM [13]	91.38	-	-
GF-SVM [17]	-	2.5	0.31
FWP-SVM-GA-1 (SVM performance of the optimal feature subset is selected)	96.61	0.07	3.39
FWP-SVM-GA-2 (SVM performance in optimizing feature weights and SVM parameters)	100	0.07	0

The false positive rate (FPR) refers to the proportion of samples that the classifier incorrectly predicts in all samples for which the actual category is negative, and the calculation

formula is given as follows:

$$\text{False positive rate (FPR)} = \frac{FP}{FP + TN} \tag{12}$$

The false negative rate (FNR) refers to the proportion of samples that the classifier incorrectly predicts in all samples for which the actual category is positive, and the calculation formula is given as follows:

$$\text{False negative rate (FNR)} = \frac{FN}{TP + FN} \tag{13}$$

The detection rate (DR) refers to the proportion of samples that the classifier correctly classifies in all samples for which the predict category is positive, and the calculation formula is given as follows:

$$\text{Detection rate (DR)} = \frac{TP}{TP + FP} \tag{14}$$

VI. CONCLUSION

This paper proposes an alarm intrusion detection algorithm (FWP-SVM-GA) based on the genetic algorithm (GA) and support vector machine (SVM) algorithm for use in a human-centered smart IDS. First, this paper makes effective use of the GA population search strategy and the capability of information exchange between individuals by optimizing the crossover probability and mutation probability of GA. The convergence of the algorithm is accelerated, and the training speed of the SVM is improved. A new fitness function is proposed that can decrease the SVM error rate and increase the true positive rate. Finally, the kernel parameter γ, the penalty parameter C and the feature weights are optimized simultaneously, and the accuracy of SVM is improved. Simulation and experimental results show that the improved intrusion detection technology based on the genetic algorithm (GA) and support vector machine (SVM) proposed in this paper increases the intrusion detection rate, accuracy rate and true positive rate; decreases the false positive rate; and reduces the SVM training time.

REFERENCES

[1] A. Sultana and M. A. Jabbar, "Intelligent network intrusion detection system using data mining techniques," in *Proc. IEEE 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Jul. 2016, pp. 329–333.

[2] L. Oneto, F. Bisio, E. Cambria, and D. Anguita, "Statistical learning theory and ELM for big social data analysis," *IEEE Comput. Intell. Mag.*, vol. 11, no. 3, pp. 45–55, Aug. 2016.

[3] A. Chaudhary, V. Tiwari, and A. Kumar, "A novel intrusion detection system for ad hoc flooding attack using fuzzy logic in mobile ad hoc networks," in *Proc. IEEE Recent Adv. Innov. Eng. (ICRAIE)*, May 2014, pp. 1–4.

[4] S. Malhotra, V. Bali, and K. K. Paliwal, "Genetic programming and k-nearest neighbour classifier based intrusion detection model," in *Proc. IEEE 7th Int. Conf. Cloud Comput., Data Sci. Amp, Eng. Conf.*, Jan. 2017, pp. 42–46.

[5] R. Sen, M. Chattopadhyay, and N. Sen, "An efficient approach to develop an intrusion detection system based on multi layer backpropagation neural network algorithm: IDS using BPNN algorithm," in *Proc. ACM SIGMIS Conf. Comput. People Res.*, 2015, pp. 105–108.

[6] M. Tabatabaefar, M. Miriestahbanati, and J.-C. Grégoire, "Network intrusion detection through artificial immune system," in *Proc. Annu. IEEE Int. Syst. Conf. (SysCon)*, Apr. 2017, pp. 1–6.

[7] T. Mehmood and H. B. M. Rais, "SVM for network anomaly detection using ACO feature subset," in *Proc. IEEE Int. Symp. Math. Sci. Comput. Res. (iSMSC)*, May 2015, pp. 121–126.

[8] N. Chand, P. Mishra, C. R. Krishna, E. S. Pilli, and M. C. Govil, "A comparative analysis of SVM and its stacking with other classification algorithm for intrusion detection," in *Proc. IEEE Int. Conf. Adv. Comput., Commun., Amp; Autom. (ICACCA) (Spring)*, Apr. 2016, pp. 1–6.

[9] R. Vijayanand, D. Devaraj, and B. Kannapiran, "Support vector machine based intrusion detection system with reduced input features for advanced metering infrastructure of smart grid," in *Proc. IEEE 4th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Jan. 2017, pp. 1–7.

[10] Q. Yang, H. Fu, and T. Zhu, "An optimization method for parameters of SVM in network intrusion detection system," in *Proc. IEEE Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, May 2016, pp. 136–142.

[11] M. S. Pervez and D. M. Farid, "Feature selection and intrusion classification in NSL-KDD cup 99 dataset employing SVMs," in *Proc. IEEE 8th Int. Conf. Softw., Knowl., Inf. Manage. Appl. (SKIMA)*, Dec. 2014, pp. 1–6.

[12] Y. Guang and N. Min, "Anomaly intrusion detection based on wavelet kernel LS-SVM," in *Proc. IEEE 3rd Int. Conf. Comput. Sci. Netw. Technol. (ICCSNT)*, Oct. 2013, pp. 434–437.

[13] T. Yerong, S. Sai, X. Ke, and L. Zhe, "Intrusion detection based on support vector machine using heuristic genetic algorithm," in *Proc. IEEE 4th Int. Conf. Commun. Syst. Netw. Technol. (CSNT)*, Apr. 2014, pp. 681–684.

[14] Z. Chen, T. Lin, N. Tang, and X. Xia, "A parallel genetic algorithm based feature selection and parameter optimization for support vector machine," *Sci. Program.*, vol. 2016, Jun. 2016, Art. no. 1.

[15] K. S. Desale and R. Ade, "Genetic algorithm based feature selection approach for effective intrusion detection system," in *Proc. IEEE Int. Conf. Comput. Commun. Inform. (ICCCI)*, Jan. 2015, pp. 1–6.

[16] W. Feng, Q. Zhang, G. Hu, and J. X. Huang, "Mining network data for intrusion detection through combining SVMs with ant colony networks," *Future Generat. Comput. Syst.*, vol. 37, pp. 127–140, Jul. 2014.

[17] H. Gharaee and H. Hosseinvand, "A new feature selection ids based on genetic algorithm and SVM," in *Proc. IEEE 8th Int. Symp. Telecommun. (IST)*, Sep. 2016, pp. 139–144.

PEIYING TAO was born in Jiangsu province. She received the B.S degree from the Nanjing University of Posts and Telecommunications in 2016, where she is currently pursuing the master's degree with the College of Computer Science. Her research interests include software defined network and big data network.

ZHE SUN was born in Shandong, China, in 1982. He received the B.S. degree from the Shenyang University of Technology in 2003, the M.S. degree from Jiangsu University in 2010, and the Ph.D. degree from Zhejiang University in 2015. He is currently with the Nanjing University of Posts and Telecommunications. His research interests include evolution computation, intelligent control, and neural network.

ZHIXIN SUN was born in Xuancheng, China, in 1964. He received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1998. From 2001 to 2002, he held a Postdoctoral position with the School of Engineering, Seoul National University, South Korea. He is currently a Professor and the Dean of the School of Modern Posts, Nanjing University of Posts and Telecommunications. His research interests are in cloud computing, cryptography, and traffic identification.