A hybrid feature selection algorithm combining information gain and genetic search for intrusion detection

Yingshang Liu\(^1\), Shouyu Liang\(^2\), Wenchong Fang\(^2\), Zhifeng Zhou\(^2\), Rong Hu\(^1\), Huafeng Zhou\(^2\), Jian Hou\(^3\), Yichang Wang\(^4\)

\(^1\)China Southern Power Grid
\(^2\)China Southern Power Grid Dispatching Power Dispatching Control Center
\(^3\)Nanning Power Supply Bureau, Guangxi Power Grid
\(^4\)Zhaotong Power Supply Bureau, Yunnan Power Grid

shcorh@163.com

Abstract. Network attacks are one of the main threats to the stable operation of smart grid equipment. As a real-time monitoring system to prevent network attacks, intrusion detection is widely used in smart grid protection. However, the massive data in the network transmission process contains a large number of redundant and irrelevant features, which makes it difficult for the intrusion detection system to process in time and reduce the efficiency. Feature selection is a method to solve this kind of problem. It can improve the speed of intrusion detection by filtering the characteristics of massive data. Therefore, a hybrid feature selection algorithm which combines information gain and genetic search to improve the work efficiency of intrusion detection systems is proposed. The algorithm is mainly divided into three parts. Firstly, the information gain value of all features is calculated by using information gain, according to which all features are ordered, and the ordered features is ranked according to an exponential increase strategy; secondly, the ranked features is used to guide the genetic algorithm search process, and a new fitness function can be used to control the search direction of genetic algorithm; finally, a classification algorithm is used to test the dataset after feature selection. In experiments, by comparing with other feature selection algorithms on 5 sets of high-dimensional UCI datasets, it is concluded that the IGExpGA proposed in this paper significantly improves the detection rate and detection speed. More importantly, in the KDD1998 network data, the algorithm proposed improves the detection rate to 98.8%, which is significantly better than other algorithms.

1. Introduction

Smart grid [1] is an interdisciplinary discipline that combines electrical engineering, information technology, and communication networks. The introduction of information technology has greatly improved the performance and reliability of power systems[2]. However, smart grid devices expose the grid system to the threat of network attacks, especially metering devices interacting with users[3]. Therefore, Some scholars[4] proposed to deploy an Intrusion Detection System (IDS) in the smart grid to detect network attacks.

The requirements for intrusion detection systems are firstly the detection rate and secondly the real-time nature. Only the detection speed is fast, the massive data transmitted in the network can be processed in a timely manner[5]. However, the detection rate and detection speed of intrusion detection...
systems significantly decrease, when massive data with a large number of redundant and irrelevant features need to be processed[6]. Therefore, Some scholars [7] use the feature selection algorithm to filter the massive data features in order to improve the working efficiency of the intrusion detection system.

Feature selection[8] is a process of reducing the dimension of the feature space by selecting some of the most effective features from the feature set, and can be divided into filtering, wrapper, and hybrid three classes, according that whether it uses the classification algorithm[9]. The filtering algorithm[10] generally evaluate features by directly using the statistical performance of the dataset, which has nothing to do with the classification algorithm and is fast. However, the filtered feature subsets have poor recognition, and the strategies of deleting feature are blind by using filtering algorithm[11]. The wrapper algorithm[12] usually search the feature set by using a search algorithm, and the searched feature subset is evaluated by the accuracy of classification algorithm, which makes the screening efficiency of feature subset is low. In additional, the genetic algorithm is easy to fall into a local optimal solution[13]. For the shortcomings of filtering and wrapper, hybrid algorithms [14] have been proposed, the main idea of which is that use the wrapper algorithm more accurately screen the features obtained by the preliminary filtering algorithm. However, it is not completely overcome the shortcomings of filtering and wrapper algorithms.

To solve above problem, this paper proposes a hybrid feature selection algorithm combining information gain and genetic search for improving the efficiency of the intrusion detection system. Its main process as is follows: The information gain value of all features is calculated using a filter algorithm based on information gain, according to which all features are ordered, and the ordered features are ranked according to an exponential increasing strategy; The ranked features are passed to the wrapped algorithm based on genetic search, which is used to guide the search of the genetic algorithm to avoid the genetic algorithm falling into the local optimal solution; The C4.5 algorithm regarded as the evaluation algorithm is used to evaluated the feature subset searched by wrapped algorithm. In the experiment, compared with the traditional UCI data set feature selection algorithm, the effectiveness of IGExpGA algorithm is verified. The IGExpGA proposed was applied to the KDDCUP1998 dataset to achieve Application in intrusion detection as well as.

2. Related work

Feature selection is a common method to improve the working efficiency of intrusion detection systems by screening valuable features from massive network dat[a8][12]. According to whether the process of feature selection uses the classification algorithm, the feature selection can be divided into three classes: filter, wrapper and hybrid algorithms[9]. Therefore, this section mainly focuses on the three different feature selection algorithms in intrusion detection applications.

Filtering algorithms evaluate feature mainly based on the statistical performance of the dataset, and are widely used in intrusion detection systems because of their fast feature selection speed. Azhagusundari [15]et al. proposed a filtering algorithm based on information gain to solve the problem of redundant features in network data. It calculates the information gain of each feature and screens feature set by setting a threshold. Osanaiye [16] et al. proposed that combining IG, GR, and ReliefF select the combined features of network data, and a decision tree is used for detecting screened feature set. The results show that it effectively reduces the redundant features of network data and significantly improves the detection rate of intrusion detection systems. Considering the correlation between each feature of network data, Ambusaidi[6]proposed selecting feature set by calculating mutual information of features and detecting the selected feature set by using a least squares support vector machine. It has higher detection rate and lower computational cost on KDD Cup 99 dataset.

The wrapped algorithm selects features based on search and evaluation algorithms. Sindhu[17]et al. combined multiple decision tree algorithms as evaluation algorithms based on genetic search. Some scholars [25]have proposed a wrapped algorithm based on multi-objective genetic search (NSGA) which aims to solve problem that genetic search is easy to fall into the local optimal solution. The experimental results show that it can select the fewest feature combinations with the largest weight.
Raman[18] et al. proposed a wrapped algorithm based on hypergraph genetic search (HG-GA) which uses the Hypergraph strategy to generate an initial population to avoid getting stuck in a local optimal solution. In addition, HG-GA also uses weighted objective functions to improve the detection rate and false alarm rate.

The hybrid algorithm will combine the advantages of filtering and wrapped algorithm. Bostani[19] et al. proposed a hybrid feature selection algorithm combining mutual information and binary gravity search (MI-BGSA), which uses mutual information to initially filter features and uses binary-based gravity search algorithm accurately selects the filtered features. The effectiveness of MI-BGSA has been proved on the NSL-KDD dataset. Gu[20] et al. proposed a hybrid algorithm for intrusion detection systems. It orders features by using a classifier at the filtering stage, and then search for features by using the Sequential Forward Selection (SFS) algorithm. It can get accurate detection rates in multiple network data. Some scholars[21] have proposed combining mutual information with genetic algorithms, which uses mutual information to filter redundant features in network data, and more precisely selects feature by using genetic algorithms in which support vector machines is regarded as evaluation algorithms.

Although the hybrid algorithm solves the shortcomings of the filtering and wrapped algorithms to a certain extent, the hybrid algorithm is only a simple combination of the reduced filtering and the wrapped methods, which uses the filtering to perform preliminary feature filtering, and then the filtered features are transmitted to the wrapped for feature selection. There are some problems with this method, such as the threshold setting of the filtering stage is still blind, and the search algorithm of the wrapped stage is easy to fall into the local optimal solution. More importantly, this hybrid approach does not achieve true convergence.

3. A Hybrid Feature Selection Algorithm Fusion of Information Gain and Genetic Algorithm

3.1. Feature ranking

Information gain[15] is a common filtering algorithm. The larger the information gain of a feature in a sample, the larger the amount of information it contains. In filtering feature selection, by calculating the information gain value contained in each feature, the feature with the highest information gain value is the highest discrimination in a given feature set. Then the information gain of feature F can be defined as:

\[ IG(F) = -\sum_{i=1}^{m} P(C_i) \log P(C_i) + P(F) \sum_{i=1}^{m} P(C_i|F) \log P(C_i|F) + P(\bar{F}) \sum_{i=1}^{m} P(C_i|F) \log P(C_i|\bar{F}) \]  

(1)

where, \( P(C_i) \) represents the probability of the category feature \( C_i \) appearing; \( P(F) \) represents the probability of the condition feature \( F \) appearing; \( P(C_i|F) \) represents the probability of the category feature \( C_i \) appearing after the condition feature \( F \) appearing; accordingly, \( P(\bar{F}) \) represents the probability that the conditional feature \( F \) does not appear; and \( P(C_i|\bar{F}) \) represents the probability that the category feature does not appear when the conditional feature \( F \) does not appear as well as.

The traditional filtering algorithm calculates the information gain of each feature according to which the features are ordered, then feature subset is output by setting the threshold to delete the feature set whose information gain value is lower than the threshold. However, the setting of the threshold is mostly based on experience and has certain blindness. Therefore, this paper proposes a filtering algorithm based on exponential ordering of information gain.

Let the initial feature set \( F_j = \{ \} \), and then calculate the information gain value of each feature according to equation (1). The features are ordered according to the information gain value, and \( IG_{F_k} = \{IG_{F_1}, IG_{F_2}, \cdots, IG_{F_n}\} \) is obtained. Then, the exponentially increasing ranking function can be defined as follows:

\[ F_{j+1} = F_j + IG_{F_1} \ast \exp(1 - t/T_{max}) \]  

(2)

where, \( t \) is the current number of iterations, and \( T_{max} \) is the maximum number of iterations.
From equation (2), the exponential incremental strategy proposed in this paper searches all features as much as possible, rather than blindly deleting features below the threshold. Then, the description of the information gain ranking algorithm based on exponentially increasing is as follows:

Algorithm 1 Information gain ranking algorithm based on exponentially increasing

Input: Dataset S, Feature set F
Output: Exponentially ranked hierarchical feature set $F'$

Initialize the feature set $I_G F$

For all features $F$ do

Calculate information gain of each feature according to equation (1)

Return feature set $I_G F$ calculated information gain

End For

For all information gain $I_G F$ do

If $I_G F_i < I_G F_{i+1}$ do

Exchange their indexes

End If

Return feature set $I_G F$ ordering according to information gain

End For

For $I_G F$ do

Exponential to rank the ordered feature set $I_G F$ according to equation (2)

End For

Return ranked feature set $F_{j+1}$

3.2. Feature Coding

In feature selection, there are only two cases where features are selected or not selected. Therefore, binary coding is required, that is, each binary bit corresponds to a feature in the feature set. A binary code value of "1" indicates that the feature is selected, and a "0" indicates that the feature is not selected. Different from the traditional encoding method, this paper adopts a hierarchical encoding strategy, and encodes each exponentially increasing interval to form k individuals $h = \{h_1, h_2, \cdots, h_k\}$, where, the total length of k individuals is N (the number of feature sets).

3.3. Fitness Function

In the search process of the genetic algorithm, the fitness function is the sole basis for controlling the search direction. Traditional fitness functions generally use classification algorithms to test selected feature sets and evaluate them based on the accuracy of the classifier. However, this strategy only considers the classification accuracy of the feature subset, and does not consider the number of feature subsets. Therefore, this paper proposes a fitness function that integrates classification accuracy and the number of feature subsets. The specific definitions are as follows:

For each selected individual $h = h_1, h_2, \cdots, h_k$, the dataset is divided into training set and test set, the training set is used to train the classifier, and then the test is performed on the test set. The support vector machine classification is calculated precision. For the number of selected feature subsets (Size($h_i$)), the ratio of all the feature subsets to Size($h_i$) is calculated. Then, the new fitness function can be defined as:

$$F(h_i) = (1 - \lambda) Accuracy(h_i) - \lambda \frac{\text{Size}(h_i)}{\text{Size}(h_i)}, i = 1, 2, \cdots, h_k, j = 1, 2, \cdots, N$$

(3)

where, $\lambda$ is a weighting factor, which is a random number between (0,1) used to balance the classification accuracy.

3.4. Individual Evolution

The evolution process of genetic algorithms is mainly divided into individual selection, crossover and mutation. Roulette is a common individual selection strategy. The basic principle is that features with
high fitness are easier to select. Assuming the chromosome size is m, and using G(x) to represent the fitness of an individual \(x_j\) as \(F(x_j)\), then the probability that it will be selected is:

\[
P(h_j) = \frac{f(h_j)}{\sum_{j=1}^{N} f(h_j)}, j = 1, 2, \ldots, N
\]  

where, N is the population size. From equation (4), it can be known that if the fitness of individual \(h_i\) is higher, the proportion of it in the overall population will be larger, and the individual \(h_i\) will be more easily inherited to the next generation.

Individual crossing can generate new individual combinations. The positions that need to be crossed are selected and interchanged according to the crossing probability \(p_c\). The individual mutation can ensure the diversity of the population, avoiding the genetic algorithm to fall into the local optimal solution. The position that needs to be mutated by the mutation probability \(p_m\) is selected. The process of mutation is that "1" becomes "0" or "0" becomes "1" in the corresponding position.

**Input:** Exponentially ranked feature set

**Output:** Optimal individual combination (feature subset)

Initialize N individuals with length l, the number of exponential groups i

For \(Gen_{\text{out}} \ll \text{Max}_k\) do \(\text{ \text{The maximum number of iterations outside the group is Max}_k}\)

For \(Gen_{\text{in}} \ll \text{Max}_l\) do \(\text{ \text{The maximum number of iterations in the group is Max}_l}\)

An initial population is generated within the feature group, that is, N individuals (each individual is a potential optimal subset of features)

Perform individual selection, individual crossover and individual mutation operations

Calculate the individual fitness function within the group according to equation (3)

Update individual \(p\) according to fitness function

End For

Calculate the fitness function of individuals outside the group according to equation (4)

Update the individual subset according to the fitness function

End For

Return Optimal individual combination

4. Experimental results and analysis

4.1. Experiment Setting

In order to verify the performance of the IGExpGA proposed in this paper, we selected 5 high-dimensional UCI datasets for testing. These five datasets come from different fields, with a maximum of 765 features. In addition, we selected the KDDCUP1998 dataset for the intrusion detection system, which provides network connection data collected from an analog local area network. Each sample has 41 features and is labeled with category information (Normal, Dos, U2r, R21, Probe). Due to the large number of KDDCUP1998 datasets, we selected 10% of them for experiments. In addition, KDDCUP1998 includes continuous and discrete data. Before feature selection, continuous data is first discretized. The detailed information of all datasets is shown in Table 1:

**Table 1. The Information of Dataset**

| Dataset        | Number of Features | Number of Samples | Number of classes | Data Type       |
|----------------|--------------------|-------------------|-------------------|-----------------|
| eighthr        | 72                 | 4736              | 2                 | UCI Dataset     |
| Comma          | 96                 | 3916              | 2                 | UCI Dataset     |
| Pd             | 754                | 756               | 2                 | UCI Dataset     |
| Arrhythmia     | 279                | 452               | 2                 | UCI Dataset     |
| Sonar          | 60                 | 208               | 2                 | UCI Dataset     |
| KDDCUP1998     | 42                 | 104857            | 5                 | Network Dataset |

5
The quality of the feature selection algorithm is usually tested by the classifier on the dataset after feature selection, and the feature selection algorithm is evaluated by recording the performance indicators of the classifier. This paper selects the detection rate and false alarm rate as the evaluation indicators to evaluate the quality of the classification algorithm. In addition, the number of feature subsets of all datasets after feature selection is recorded for assessing detection speed. The experiment is divided into two parts: Firstly, it can verify the effectiveness of the IGExpGA algorithm to compare with the traditional feature selection algorithm on 5 sets of high-dimensional UCI datasets. Secondly, the IGExpGA algorithm is applied to intrusion detection system to verify practical value on the KDDCUP1998 dataset.

4.2. Validation on UCI dataset

4.2.1 The feature selection of IGExpGA

From the working principle of IGExpGA algorithm, we can first use the information gain algorithm to calculate the information gain of all features, and order the features according to the information gain value; then group the ordered feature sets according to the exponential increase strategy; finally the genetic algorithm is used to perform hierarchical evolution on the hierarchical feature set. Fig 1 shows the feature selection process of the IGExpGA algorithm.

It can be seen from Fig.1 that with the increase of the feature index level, the TP rate and FP rate of the C4.5 algorithm on the five UCI datasets show different trends. Form Fig.1 (a), it is found that the detection rate increases first and then decreases as the feature index level increases, and the highest detection rate is often obtained in the middle. This is mainly based on two principles: 1) At the beginning, the number of candidate features is large and contains a large number of redundant features; 2) At the end, the number of candidate features is small, which makes the genetic algorithm fall into a local optimum during the search solution.

4.2.2 Comparison with traditional feature selection algorithms

In order to verify the effectiveness of the IGExpGA proposed in this article, this section selects eight traditional feature selection algorithms. The filtering algorithms are information gain (IG), weight distance (ReliefF), and correlation measure (Chis). The wrapped algorithms are genetic search (GA), particle swarm search (PSO), and differential evolution search (EA). The hybrid algorithm is the fusion of information gain and genetic search (IGGA). The C4.5 algorithm was used to test the dataset after feature selection. The experimental results are shown in Table 2 and Table 3.

Table 2. Detection rate of the C4.5 algorithm on the 5 UCI datasets filtered by the feature selection algorithm

| Dataset | Filter | Wrapper | Hybrid |
|---------|--------|---------|--------|
| original | IG     | RF      | Chis  |
|         |        |         | GA    |
|         |        |         | PSO   |
|         |        |         | EA    |
|         |        |         | IGGA  |
|         |        |         | IGExpGA |
Table 3: False positive rates of the C4.5 algorithm on the 5 UCI datasets filtered by the feature selection algorithm

| Dataset       | Filter Wrappers | Hybrid Wrappers |
|---------------|-----------------|-----------------|
|               |                 |                 |
| Eighthr       | 9.5             | 7.1             |
| Comma         | 10.7            | 8.7             |
| Pd            | 25.5            | 26.0            |
| Arrhythmia    | 20.8            | 21.5            |
| Sonar         | 28.8            | 21.6            |
| Average       | 19.06           | 17.60           |

As can be seen from Tables 2 and 3, the detection rate and false positive rate of the C4.5 algorithm on the original UCI dataset are very poor, especially the detection rate on the Sonar dataset is only 71.2%, while the false positive rate was as high as 28.8%. Comparing the three filtering algorithms, the detection rate and false alarm rate of IG on the five UCI datasets are 83.84% and 18.63%, respectively, which is the best of the three algorithms. Comparing the three hybrid algorithms, it is found that the detection rate of PSO on all datasets is the highest, and the false alarm rate is also the lowest.

Compared with the hybrid algorithm, it is found that the screening effect of IGGA on the Comma dataset is better than IGExpGA, while is significantly lower on other datasets. Comparing the three feature selection algorithms, it is found that the wrapped algorithm is better than the filtering algorithm as a whole, while the average detection rate of the dataset filtered by the traditional hybrid algorithm is only 85.08%, which has a small improvement effect. Therefore, the IGExpGA proposed in this paper has the best screening effect among all feature selection algorithms.

In addition, detection speed is one of the methods to evaluate the effectiveness of feature selection algorithms. The number of filtered feature subsets can reflect the detection speed of the algorithm. Therefore, this paper gives the number of feature subsets screened by eight feature selection algorithms. The specific information is shown in Table 4:

Table 4 The number of feature subsets of the 5 UCI datasets filtered by the feature selection algorithm

| Dataset       | Filter Wrappers | Hybrid Wrappers |
|---------------|-----------------|-----------------|
|               |                 |                 |
| Eighthr       | 72              | 18              |
| Comma         | 96              | 37              |
| Pd            | 754             | 245             |
| Arrhythmia    | 279             | 75              |
| Sonar         | 60              | 8               |
| Average       | 252.2           | 72.4            |

As can be seen from Table 4, the number of features in the dataset filtered by the eight feature selection algorithms has been reduced. Comparing the three filtering algorithms, it is found that the number of features filtered by Chis is the smallest, and the average number on the five UCI datasets is 118.2. Comparing the three wrapped algorithms, it is found that different results are presented on different datasets. The average number of features after EA screening is only 114, which is slightly lower than the other two algorithms. Comparing with the hybrid algorithm, it is found that the number of features filtered by the traditional hybrid algorithm is far less than that of the filtering and wrapped algorithms, and the average number of features is only 72.4. Comparing the two hybrid algorithms, it is found that the number of feature subsets screened by IGExpGA proposed in this paper is the
smallest of all feature selection algorithms (16.6). It can be seen that the IGExpGA screened dataset presented in this paper has a better detection speed.

4.3 Application on KDDCUP1998 dataset
In order to verify the feature selection effect of the IGExpGA algorithm on the KDD1998 dataset, 8 traditional feature selection algorithms are selected. The eight feature selection algorithms are the same as in the experiments in Section 4.3.2, and the C4.5 algorithm is used to test the feature subset after feature selection. Table 5 and Table 6 show the comparison results of the nine feature selection algorithms on the KDD1998 dataset.

Table 5: Detection rate and false positive rate of the C4.5 algorithm on the KDDCUP1998 data set filtered by the feature selection algorithm

| Index | Filter | Wrapper | Hybrid |
|-------|--------|---------|--------|
| TPR   | original | IG | RF | Chis | GA | PSO | EA | IGGA | IGExpGA |
|       | 97.7 | 97.5 | 97.3 | 97.1 | 98.2 | 98.1 | 97.8 | 98.4 | 99.7 |
| FPR   | 1.9 | 2.1 | 2.2 | 2.3 | 1.7 | 1.8 | 2.0 | 1.5 | 0.2 |

As can be seen from Table 5, the C4.5 algorithm has a poor classification effect on the original KDDCUP1998 dataset, with detection rates and false positive rates of 98.5% and 1.3%, respectively. Comparing the three filtering algorithms, it is found that, as the same as in the previous experimental results, the dataset after IG feature selection has a higher detection rate. GA is the best feature selection method in the wrapped algorithm. Compared with the hybrid algorithm, IGExpGA not only has a higher detection rate, but also has the lowest false positive rate, which significantly improves the efficiency of intrusion detection system. We compared the number of feature subsets screened by the eight feature selection algorithms. The experimental results are shown in Table 6.

Table 6: Number of feature subsets of the KDDCUP1998 data set filtered by the feature selection algorithm

| Index   | Filter | Wrapper | Hybrid |
|---------|--------|---------|--------|
| KDDCUP1998 | original | IG | RF | Chis | GA | PSO | EA | IGGA | IGExpGA |
|         | 42 | 24 | 26 | 25 | 20 | 21 | 22 | 8 | 6 |

It can be known from Table 6 that the number of feature subsets filtered by the eight feature selections is reduced. Unlike previous experiments, IG has the smallest number of feature subsets among the three filtering algorithms, while the GA has the smallest number of feature subsets among the three wrapped algorithms. As the same as the previous experimental results, the data set filtered by the hybrid algorithm has the smallest number of feature subsets. The number of feature subsets screened by the traditional hybrid algorithm (IGGA) is only 10, while the IGExpGA proposed in this paper has the smallest number of feature subsets (6). Therefore, the IGExpGA screened data set presented in this paper has a better detection speed.

From above results, it is found that the detection rate and false alarm rate of the data set filtered by the IGExpGA algorithm proposed are optimal, and the number of feature subsets after screening is the smallest. In order to observe which features in the feature subset selected by the eight feature selection algorithms are most advantageous for classification. The specific feature subsets screened by eight feature selection algorithms are shown in Table 7.

Table 7: Feature subsets filtered by 8 feature selection algorithms

| Algorithms | Feature subsets |
|------------|----------------|
| IG         | 3,5,23,6,12,36,24,32,37,4,34,33,30,2,38,25,35,29,39,26,31,40,27,41,28,11,10 |
| RF         | 2,3,4,5,8,10,12,17,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41 |
| Chis       | 5,3,6,23,37,35,12,4,33,30,32,34,36,18,25,24,29,11,38,39,2,26,27,40,31,41,28,14,10 |
| GA         | 2,3,4,5,6,7,8,10,11,17,18,20,21,23,24,26,27,30,33,34,35,38 |
As can be seen from Table 8, the number of feature subsets filtered by the three filtering algorithms is the largest. Since the filtering algorithm sorts the filtered feature subset, we find that only the third feature (the type of network service of the target host) are selected in the top 10 features of the three filtering algorithms. Observing the three wrapped algorithms found that the third feature was also selected. In addition, the three wrapped algorithms also select features such as Nos. 5, 23, and 24. Therefore, it can be concluded that the feature subset screened by the IGExpGA algorithm is crucial for intrusion detection.

5. Conclusion

Intrusion detection is an effective measure to prevent smart grid equipment from being attacked. However, the massive network data contains a lot of redundant and irrelevant features, which leads to a significant decrease in the detection rate and detection speed of the intrusion detection system. Therefore, this paper proposes a hybrid feature selection algorithm combines information gain and genetic search to improve the efficiency of the intrusion detection system by selecting the feature of massive network data. The algorithm is mainly divided into three parts: Firstly, to solve the blind setting threshold problem of the traditional filtering algorithm, an exponential information gain increasing strategy is proposed in this paper, which is used to exponentially increase the feature set according to the information gain. Secondly, in view of the "premature" problem of genetic algorithms in traditional packaged algorithms, this paper proposes a hierarchical evolution strategy to guide the search process of genetic algorithms. In addition, a fitness function that integrates the classification accuracy and scale of feature subsets is proposed. Finally, a support vector machine is used to test the dataset after feature selection. The experimental results show that the detection rate and speed of the IGExpGA proposed in this paper on the five high-dimensional UCI datasets are significantly better than the traditional feature selection algorithms. More importantly, the detection rate of the support vector machine on the KDD1998 data set filtered by IGExpGA mentioned is as high as 99.8%, which significantly improves the working efficiency of the intrusion detection system.

References

[1]. Dileep G. A survey on smart grid technologies and applications[J]. Renew. Energy, 2020, 146: 2589-2625.
[2]. Fang X, Misra S, Xue G, et al. Managing smart grid information in the cloud: opportunities, model, and applications[J]. IEEE network, 2012, 26(4): 32-38.
[3]. Gai K, Qiu M, Ming Z, et al. Spoofing-jamming attack strategy using optimal power distributions in wireless smart grid networks[J]. IEEE Transactions on Smart Grid, 2017, 8(5): 2431-2439.
[4]. Patel A, Alhussian H, Pedersen J M, et al. A nifty collaborative intrusion detection and prevention architecture for smart grid ecosystems[J]. Computers & Security, 2017, 64: 92-109.
[5]. Al-Yaseen W L, Othman Z A, Nazri M Z A. Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system[J]. Expert Systems with Applications, 2017, 67: 296-303.
[6]. Ambusaidi M A, He X, Nanda P, et al. Building an intrusion detection system using a filter-based feature selection algorithm[J]. IEEE transactions on computers, 2016, 65(10): 2986-2998.
[7]. Ambusaidi M A, He X, Nanda P, et al. Building an intrusion detection system using a filter-based feature selection algorithm[J]. IEEE transactions on computers, 2016, 65(10): 2986-2998.
[8]. Xue B, Zhang M, Browne W N, et al. A survey on evolutionary computation approaches to feature selection[J]. IEEE Transactions on Evolutionary Computation, 2015, 20(4): 606-626.
[9]. Lee S J, Xu Z, Li T, et al. A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making[J]. Journal of biomedical informatics, 2018, 78: 144-155.

[10].Lazar C, Taminau J, Meganck S, et al. A survey on filter techniques for feature selection in gene expression microarray analysis[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(4): 1106-1119.

[11].Ghosh M, Guha R, Sarkar R, et al. A wrapper-filter feature selection technique based on ant colony optimization[J]. Neural Computing and Applications, 2019: 1-19.

[12].Kabir M M, Islam M M, Murase K. A new wrapper feature selection approach using neural network[J]. Neurocomputing, 2010, 73(16-18): 3273-3283.

[13].Ghareb A S, Bakar A A, Hamdan A R. Hybrid feature selection based on enhanced genetic algorithm for text categorization[J]. Expert Systems with Applications, 2016, 49: 31-47.

[14].Solorio-Fernández S, Carrasco-Ochoa J A, Martínez-Trinidad J F. A new hybrid filter–wrapper feature selection method for clustering based on ranking[J]. Neurocomputing, 2016, 214: 866-880.

[15].Azhagusundari B, Thanamani A S. Feature selection based on information gain[J]. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2013, 2(2): 18-21.

[16].Osanaiye O, Cai H, Choo K K R, et al. Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing[J]. EURASIP Journal on Wireless Communications and Networking, 2016, 2016(1): 130.

[17].Sindhu S S S, Geetha S, Kannan A. Decision tree based light weight intrusion detection using a wrapper approach[J]. Expert Systems with applications, 2012, 39(1): 129-141.

[18].Raman M R G, Somu N, Kirthivasan K, et al. An efficient intrusion detection system based on hypergraph-Genetic algorithm for parameter optimization and feature selection in support vector machine[J]. Knowledge-Based Systems, 2017, 134: 1-12.

[19].Bostani H, Sheikhan M. Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems[J]. Soft computing, 2017, 21(9): 2307-2324.

[20].Gu Y, Li K, Guo Z, et al. Semi-supervised k-means ddos detection method using hybrid feature selection algorithm[J]. IEEE Access, 2019, 7: 64351-64365.

[21].Vijayanand R, Devaraj D, Kannapiran B. A novel intrusion detection system for wireless mesh network with hybrid feature selection technique based on GA and MI[J]. Journal of Intelligent & Fuzzy Systems, 2018, 34(3): 1243-1250.