A feature selection algorithm combining information gain and multi-objective genetic search for intrusion detection system

Tao Xie*

Ningxia Institute of science and technology, College of computer science and engineering, Shizuishan Ningxia 753000, China

Abstract. In order to improve the detection rate and speed of intrusion detection system, this paper proposes a feature selection algorithm. The algorithm uses information gain to rank the features in descending order, and then uses a multi-objective genetic algorithm to gradually search the ranking features to find the optimal feature combination. We classified the Kddcup98 dataset into five classes, DOS, PROBE, R2L, and U2R, and conducted numerous experiments on each class. Experimental results show that for each class of attack, the proposed algorithm can not only speed up the feature selection, but also significantly improve the detection rate of the algorithm.

1 Introduction

Intrusion detection system (IDS)[1] prevent or mitigate threats from network attacks without affecting network performance. The requirement for intrusion detection is to ensure the detection rate, detection speed is crucial. Due to access information from the online world is huge. The detection speed is too slow, which may lead to an intrusion situation. How to improve the detection speed of IDS under the premise of correct detection has become the hot spot of current research.

In recent years, studies [2] have shown that the redundancy features in access information are the main reason for the slowdown of IDS. Many researchers [1-6] select the core features that can be identified by selecting the features of access information. Therefore, by selecting features of the access information, retaining important features that can represent the access information is an effective method to improve the detection speed.

Based on the above description, this paper proposes a feature selection algorithm combining information gain and multi-objective genetic search. Firstly, the value of the conditional feature is calculated by using the information gain, and the conditional feature is sorted in descending order. Then multi-objective genetic algorithm is used to search the ranked conditional features, and two objective functions are set to find the optimal feature subset. Finally, C4.5 is used to test the data after feature selection to verify the effectiveness of the method. In experiments, compared with many traditional feature selection algorithms,

* Corresponding author: 44021984@qq.com

to verify the effectiveness of IG-MGA algorithm, and the algorithm is applied to Kddcup98 dataset.

2 Related work

In this section, we consider reducing the number of features to improve the detection efficiency. Feature selection generally includes metrics index and search strategy.

In metrics index, Ambusaidi et al [1]. use mutual information to select features of network data and use support vector machine as detection algorithm. Experimental results show that the proposed algorithm is superior to the most advanced method. Hamid et al [2]. used mutual information for preliminary feature selection, which effectively weakened the correlation between features. Then use the BGSA algorithm for more precise feature selection.

Sangkatsanee et al [3]. used information gain as a metric, and selected 12 core features by setting thresholds. Then use algorithms such as decision trees for detection. The algorithm is characterized by high speed and high efficiency to select the feature. Alazzam et al [4]. used pigeon search to search feature subsets, and the fitness function fully considered the accuracy and number of features. Experimental results show that the algorithm is superior to the most feature selection algorithms.

Vijayanand et al [5]. used genetic algorithm to conduct feature search, and used support vector machine as detection algorithm. Experimental results show that the SVM has the best effect. However, in the feature search, genetic algorithm loses population diversity and falls into a local optimal solution. Therefore, Khammassi et al [6]. used multi-objective Genetic algorithm (NSGA-II) to conduct feature search, and used logistic regression as the detection algorithm. Experiments results show the effectiveness of the proposed algorithm.

3 The proposed method

3.1 Feature ranking

When calculating the weight of features, Information Gain (IG) only considers the information of features to classes when features appear or do not appear. The amount of information is used as the weight of the feature to select the feature. Then, the IG is calculated as:

$$\mathrm{IG(F)} = \mathrm{P(F)}\sum \mathrm{P}\left(\frac{\mathrm{C_t}}{\mathrm{F}}\right)\log\frac{\mathrm{P(C_t/F)}}{\mathrm{P(C_t)}} + \mathrm{P(\overline{F})}\sum \mathrm{P}\left(\frac{\mathrm{C_t}}{\mathrm{\overline{F}}}\right)\log\frac{\mathrm{P(C_t/\overline{F})}}{\mathrm{P(C_t)}} \tag{1}$$

By calculating the information gain value of each feature, and ranking the features according to the value. The method is to set a threshold value to remove the feature below the threshold value.

3.2 Feature selection algorithm combining IG and NSGA-II

3.2.1 Feature coding

In the feature selection, the conditional feature only exists whether or not to be selected, so we use the binary encoding strategy to encode the feature. Suppose the feature $F = \{f_1, f_2, \cdots, f_m\}$, a string of length m and encoded to represent a feature combination, namely:

$$G = \{g_1, g_2, \cdots, g_m\} \tag{2}$$

where $\mathrm{g_i} = 0$ or 1, $g_i = 0$, it means that the feature corresponding to g_i is not selected in feature combination. If not, select the feature.

3.2.2 Objective function

In feature search, multi-objective genetic algorithm can set multiple objective functions. Therefore, this paper selects the accuracy of the C4.5 algorithm and the number of selected feature subsets as the objective function. Then, the objective function can be defined as:

$$\mathrm{Fit} = \begin{cases} \mathrm{min} & |\mathrm{F}| \\ \mathrm{max\,accuracy(F)} \end{cases} \quad \mathrm{s.t}\,|\mathrm{F}| \ll |\mathrm{F}|_{\mathrm{IG}} \tag{3}$$

where, F is the subset of features; $|F|$ is the number of feature subsets; $accuracy(\quad)$is the accuracy of the feature subsets; $|F|_{IG}$ is the number of feature subsets is less than that after ranking based on information gain.

3.2.3 Genetic operator

Selection operator: The binary tournament selection strategy is adopted. Two parent individuals are selected from the population with the same probability, and the two individuals are compared according to the crowding comparison operator, and the optimal individual is selected to enter the next generation population.

Crossover operator: The uniform crossover strategy is adopted. Two parent individuals are exchanged in pairs, and a certain gene in the two parent individuals is crossed according to the probability P_c to form two new individuals. The specific operation is as follows:

$$\{\mathrm{C_i} = (\mathrm{c_{i1}}, \mathrm{c_{i2}}, \cdots, \mathrm{c_{iN}}), \mathrm{C_j} = \left(\mathrm{c_{j1}}, \mathrm{c_{j2}}, \cdots, \mathrm{c_{jN}}\right) \xrightarrow{\mathrm{P_c}} \mathrm{C_i} = \left(\mathrm{c_{i1}}, \mathrm{c_{j2}}, \mathrm{c_{i3}}, \mathrm{c_{j4}} \cdots, \mathrm{c_{jN}}\right), \mathrm{C_j} = \left(\mathrm{c_{j1}}, \mathrm{c_{i2}}, \mathrm{c_{j3}}, \mathrm{c_{i4}} \cdots, \mathrm{c_{iN}}\right) \tag{4}$$

Mutation operator: The reversal mutation strategy is adopted. One gene of the parent individuals is flipped, and two of the parent individuals are selected to mutation according to the probability P_m, so as to form a new individual. The specific operation is as follows:

$$\mathrm{C_i} = (\mathrm{c_{i1}}, \mathrm{c_{i2}}, \cdots, \mathrm{c_{iN}}) \xrightarrow{\mathrm{P_m}} (\mathrm{c_{i1}}, \mathrm{c_{iN-1}}, \cdots, \mathrm{c_{i2}}, \mathrm{c_{iN}}) \tag{5}$$

3.2.4 Steps of Algorithm

In the feature selection, multi-objective genetics uses "fast non-dominant sequencing" to Pareto decomposition of the solutions in the group. The solution with low non-dominated level is better than the solution with high non-dominated level.

If the two solutions are on the same front surface, the solution with a smaller crowding distance is better. Then, the feature selection algorithm combining feature ranking and multi-objective genetic search can be described as follows:

Input: Dataset (Including N features)
Output: Non-dominant solution F

1. Calculate The information gain value $\mathrm{IG(F_i)}$ of each feature
2. Descending sort ranking to the IG value
3. Delete features below the threshold

4. The initial group $\mathrm{P_o}$ was randomly generated and the objective function of each individual was calculated.
5. Calculate the non-dominant rank of each individual in the $\mathrm{P_o}$ using fast non-dominant rank
6. The crowding distance is used to calculate the crowding distance of each individual in the $\mathrm{P_o}$
7. **Repeat**
8. Binary elite selection, uniform crossover, and one-bit flip mutation were performed for each individual in the $\mathrm{P_o}$
9. Recalculate the objective function of each individual
10. Recalculate the non-dominant level for each individual in the $\mathrm{P_o}$
11. Recalculate the crowding distance of each individual in the $\mathrm{P_o}$
12. **Until** Trmination condition

4 Intrusion detection experiment

4.2 Experimental data

In this section, 10% of Kddcup98 dataset is selected for intrusion detection experiment. The dataset contains 494,021 samples, and each sample contains 41 features, which is a typical high-dimensional large sample dataset. We divided the data into five parts: DOS, PROBE, R21, U2R and NORMAL. Where NORMAL does not contain any attacks. Table 1 shows the number of samples and specific attack types contained in each class in the Kddcup98 dataset.

4.2 Experimental settings

Feature selection algorithm in Kddcup98 dataset how the selection effect often needs to be tested after feature selection. Therefore, we use C4.5 algorithm to test the dataset, and record the C4.5 algorithm training time and detection rate indexes. To this end, we give the definition of detection rate:

$$\mathrm{DER} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{6}$$

where, TP is the number of correct samples tested, and $TP + FN$ is all samples.

4.3 Result of feature selection

This section presents the results of many feature selection algorithms in Kddcup98 dataset. There are information gain, ReliefF and Chis in the metrics. There are genetic, particle swarm and multi-objective genetic algorithms in the search strategy. In addition, we are combined of metric and search strategy. Table 1 shows the experimental results.

From Table 1, some features in the KDDCUO98 dataset are very important. No matter which feature selection algorithm is used, the feature is always selected. For example, count is selected by all feature selection algorithms, and it represents the number of connections that have passed through the same node in the past 2 seconds. These core features helps to understand which features are carried by intrusion behaviors, and can significantly improve the effect of intrusion detection.

Table 1. Feature subsets selected by 9 feature selection algorithms.

Algorithm	Selected features
IG	1,2,3,4,5,6,12,23,24,25,26,27,29,30,31,32,33,34,35,36,37,38,39,40,41
ReliefF	2,3,4,8,12,23,24,25,26,29,30,31,32,33,34,35,36,37,38,39,40,41
Chis	1,2,4,5,6,10,12,14,23,24,25,27,29,30,31,32,33,34,35,36,37,38,40
PSO	1,2,3,4,5,6,7,9,11,12,14,15,16,20,21,22,23,24,27,30,31,32,35,36,38,39,40
GA	1,2,3,4,5,6,11,12,14,17,18,20,21,22,23,25,27,30,31,39,40
MGA	1,3,4,5,6,11,12,14,17,18,20,21,22,23,25,27,30,31,39,40
IG-GA	1,3,4,5,6,11,23,25,29,30,32,35,36,37
IG-PSO	1,2,3,4,5,6,12,23,24,30,34,36.37.39,40,41
IG-MGA	1,2,3,5,12,23,30.35.36,38,40

4.4 Effect of feature selection

This section selects the C4.5 as the detection algorithm to experiment on the Kddcup98 dataset, and records the detection rate and training time. The experimental results are shown in Table 2.

Table 2. C4.5 indexes in Kddcup98 dataset after feature selection.

	Origianl	IG	ReliefF	Chis	PSO	GA	MGA	IG-GA	IG-PSO	IG-MGA
NORMAL	100.0	99.9	99.8	99.9	100.0	100.0	100.0	100.0	100.0	99.9
DOS	99.4	99.4	98.3	99.4	99.4	99.3	99.4	99.2	99.4	99.5
PROBE	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
R21	46.2	59.6	28.3	48.1	65.4	69.2	46.2	59.6	65.4	65.5
U2R	96.4	96.2	94.3	96.4	97.5	96.8	96.4	96.6	97.5	97.2
Times/s	95.32	35.23	31.27	47.23	16.88	34.55	79.83	21.59	24.72	14.92

From Table 2, C4.5 has the worst detection rate and the longest training time in the original dataset. In the metrics index, when the attack type is U2R, the Chis algorithm has the best feature selection effect. On the whole, IG algorithm has the best effect. In the search strategy, the three algorithms have different advantages in different attack types, and GA algorithm has the best effect on the whole. In the hybrid algorithm, the feature selection effect of IG-MGA is the best. When the attack type is DOS and R21, the detection rate is 99.5% and 65.5% respectively.

5 Conclusion

This paper proposes a feature selection algorithm combining information gain and multi-objective genetic search. The experimental results on the Kddcup98 dataset show that the algorithm has the best detection rate, as well as good classification performance, which can effectively improve the security of the intrusion detection system. In addition, the C4.5 algorithm has the shortest prediction time on the feature selection Kddcup99 dataset.

References

1. M A Ambusaidi, X He, P Nanda. Building an intrusion detection system using a filter-based feature selection algorithm. IEEE transactions on computers. **65**,2986-2998(2016)
2. H Bostani, M Sheikhan. Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems. Soft computing. **21**,2307-2324(2017)
3. P Sangkatsanee, N Wattanapongsakorn. C Charnsripinyo. Practical real-time intrusion detection using machine learning approaches. Computer Communications. **34**,2227-2235.(2011)
4. H Alazzam, A Sharieh K E Sabri. A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer. Expert Systems with Applications. **148**,113249(2020)
5. R Vijayanand, D Devaraj, B Kannapiran. Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection. Computers & Security. **77**, 304-314(2018)
6. C Khammassi, S Krichen. A NSGA2-LR wrapper approach for feature selection in network intrusion detection. Computer Networks.(2020)