TELKOMNIKA, Vol. 11, No. 4, April 2013, pp. 1725~1730
ISSN: 2302-4046

Hybrid Feature Selection Based on Improved Genetic Algorithm

Shuxin ZHU[1], Bin HU[2]*
[1,2]College of Information Science and Technology, Nanjing Agricultural University
Weigang 1, Nanjing,china, 210095
*Corresponding author, email: hubin@njau.edu.cn

Abstract

High dimensionality is one of the most troublesome difficulties encountered in intrusion detection system analysis and application. For high dimension data, feature selection not only can improve the accuracy and efficiency of classification, but also discover informative subset. Combining Filter type and Wrapper type characteristics, this paper proposes a hybrid type method for feature selection using a improved genetic algorithm contained reward and punishment mechanism. The mechanism can guarantee this algorithm rapid convergence on approximate global optimal solution. According to the experimental results, this algorithm performs well and it's time complexity is low.

Keywords: *intrusion detection system, genetic algorithm (GA), feature selection, mutual information, hybrid type*

1. Introduction

Intrusion detection is essentially a classification problem [1] and the first problem to solved in classification is the feature extraction and selection. There is not a linear relationship between features and classifier performances. But when the feature number exceeds a certain limit, it will cause the classifier performance variation. The so-called feature selection [2], is to choose the output relevant or important feature subset from the original feature set according to a certain evaluation function and as far as possible to reduce the dimensionality of the feature space on the premise of not reducing the classification accuracy. Apparently, the feature selection has two key problems: selection of suitable evaluation function and efficient feature subset search methods.

According to be dependant or not dependant on machine learning algorithm, feature selection algorithms can be divided into two categories. One is wrapper-type algorithm [3] and the other is filter-type algorithm [4]. Filter based feature selection algorithm is independent of machine learning algorithm and has some benefits such as low computational cost and high efficiency, but it performs mediocrely in reducing dimension. On the contrary, wrapper based feature selection algorithms rely on one or several machine learning algorithms with the characteristic of large computation complexity, low efficiency but good effect of reducing the dimensions. This paper combines the characteristics of two kinds of algorithms and proposes a hybrid method to perform the characteristic selection in intrusion detection data set. This method utilizes mutual information method to eliminate redundancy, establishes reward punishment mechanism and uses the improved genetic algorithm to generate optimal subset suitable for intrusion detection classification.

2. Research Method

In the intrusion detection process, extraction and processing too many feature numbers is one of the main reasons leading to speed down. In fact, some parameters just include or contain minimal information on the system status and almost have no effect on the test results. These parameters are called redundancy parameters. So feature selection should first adopt methods like filter based methods to remove redundancy parameter and retain the important characteristics reflecting the state of the system. In this paper, we put forward a feature

Received November 19, 2012; Revised January 2, 2013; Accepted January 15, 2013

selection method based on mutual information from perspective of the data statistical characteristic. The use of mutual information feature selection is also based on assumption that the evaluation data and the actual data have the same statistical properties. So we can use the statistical properties of known evaluation data to express actual system characteristics.

Mutual information could measure the co-occurrence relations of feature items and categories. Characteristic xi appears in a certain category C with high probability and low probability in other categories will acquire greater mutual information value (MI). MI could be expressed as formula (1):

$$MI(x_i, C) = \log \frac{P(x_i \mid C)}{P(x_i) * P(C)} \tag{1}$$

Because MI is beneficial to the low frequency features, it easily leads to over-fitting. On the other hand, removing low-frequency features may affect the detection efficiency [5]. In order to improve it, we add the P(xi) factor function which stands for the feature appearance probability to mutual information and form an improved algorithm (PMI). PMI could be expressed as formula (2):

$$PMI(x_i, C) = P(x_i) \log \frac{P(x_i \mid C)}{P(x_i) * P(C)} \tag{2}$$

The criterion of this method is the conditional mutual information and the selection algorithm is sequential forward selection. First we let feature set empty, then calculate the current PMI (xi,c) value and choose the maximum mutual information into feature subset Fi, and in the end compute PMI (xi,c/Fi) one by one. So after choosing several features into the feature subset, a new sample space is formed. Compared with the original sample space, the new has the smaller dimension and lower correlation among feature variables.

After achieving reduction of original feature sets by eliminating the redundancy among features, we could adopt the Wrapper based feature selection method based on genetic algorithm. Figure 1 shows the improved genetic algorithm which could be used in intrusion detection system. The samples are filtered by conditional mutual information and then sent to improved genetic operator to be optimized according to classifier. In the end, the optimal feature subset is acquired.

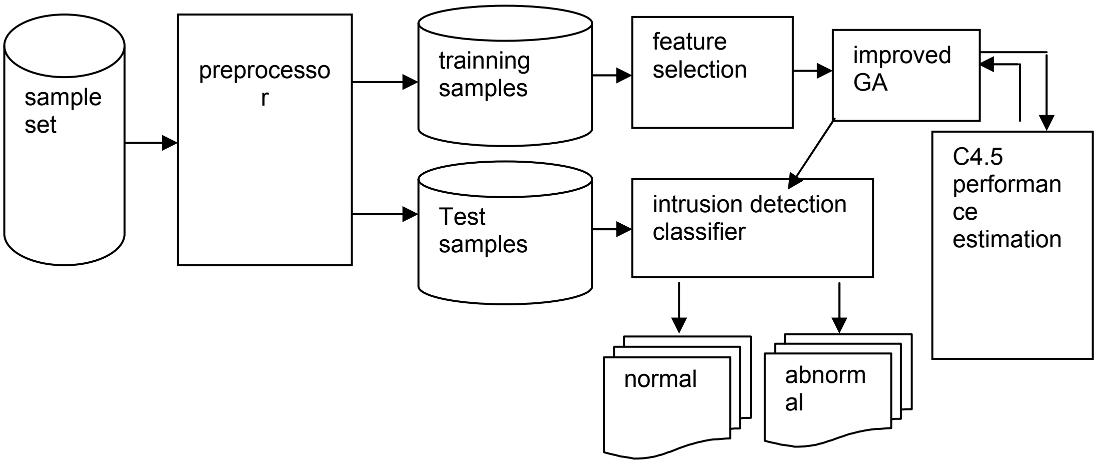

Figure 1. Feature Selection Flow Diagram in the Intrusion Detection based on Improved GA Algorithm

2.1. Improved Genetic Algorithm Design

Selecting the optimal subspace from the existing feature subset has proven to be a NP-hard problem. From the optimization point of view, the feature selection problem is a combinatorial optimization problem. The current feature selection methods based on genetic algorithm [6] usually evaluate the feature subset on the basis of classifier and give individual evaluation indexes and fitness (often expressed as f) according to the classification accuracy.

2.1.1. Coding Scheme

In the feature selection range, individual coding scheme often utilizes binary expression. Because binary string expression is simple and convenient to operate. It also can represent a wide range of different information. Assuming the original set has 16 features, then the length of individual L=16. Each individual has a gene corresponding to sequence characteristics. When the individual has a gene expressed as "1" it means that the corresponding feature item is chosen. Conversely, "0" means the feature is not used. For example, individual 1100100001011000 means that the second, fifth first, tenth, twelfth, thirteenth features are selected. If using the exhaustive search method to solve the optimal feature subset, it will be 2m possible subset combinations for the set containing m features. Such a huge search space is not feasible. By using genetic algorithm, we not only can ensure the global optimum but also avoid the huge search cost.

2.1.2. Initial Population Selection

Genetic algorithm always starts from the population which represents potential solution set of the problem, namely initial population. The population consists of several genetic individual components and each individual is a possible solution. Here we adopt the stochastic method to choose the initial group. Every gene of each individual has the same probability in {0, 1} selection. Individual size is determined according to the actual situation.

2.1.3. Design of Genetic Operator

There are three operators applied to get optimal feature subset: the proportional selection operator, single-point crossover operator, simple mutation operator.

(1) Selection is the survival of the fittest process based on fitness. Here the proportional selection operator is also called a roulette wheel selection strategy. The basic idea is that the probability of each individual’s being selected is proportional to the size of its fitness.

(2) The aim of crossover is to generate new units in the next generation. By using the crossover operation, genetic algorithm will be able to greatly improve its search capability. In this algorithm, the crossover operation uses a single-point crossover and selects individuals with a certain probability p. For two elder individuals participating in the crossover, we randomly selects one cross point and generates two new individual by swapping two individuals partial structure behind it or in front of it.

(3) Mutation is an assisted operation in genetic algorithm and it is mainly to maintain the population diversity. This operation uses the simple mutation and selects individuals to perform mutation operation in the probability of pm. It randomly selects the individual gene bit participating in the mutation and does the reverse operation.

2.1.4.Termination Condition

Because the genetic algorithm does not use of the gradient of the objective function and other information, it’s unable to use traditional methods to judge its convergence in order to terminate the genetic process. The commonly used method is by controlling the parameters to realize the algorithm termination, such as reaching the specified maximum algebra. Another way is when the average fitness-differences of adjacent generations is less than a threshold value ε, it can terminate the genetic operation and the end condition is $\left\| \left| \overline{f_{i-2}} - \overline{f_{i-1}} \right| - \left| \overline{f_{i-1}} - \overline{f_i} \right| \right\| < \varepsilon$.

2.2. Fitness Definition and Reward and Punishment Mechanism

In most wrapper based genetic algorithms feature selection methods are using certain classifier models to evaluate selected feature set and utilizing the classification contribution value or the classification error rate as the fitness function. In this paper, in order to search out the better linear separability feature subspace, we adopt C4.5 algorithm [7] as the classifier model and classification error rate as the fitness function.

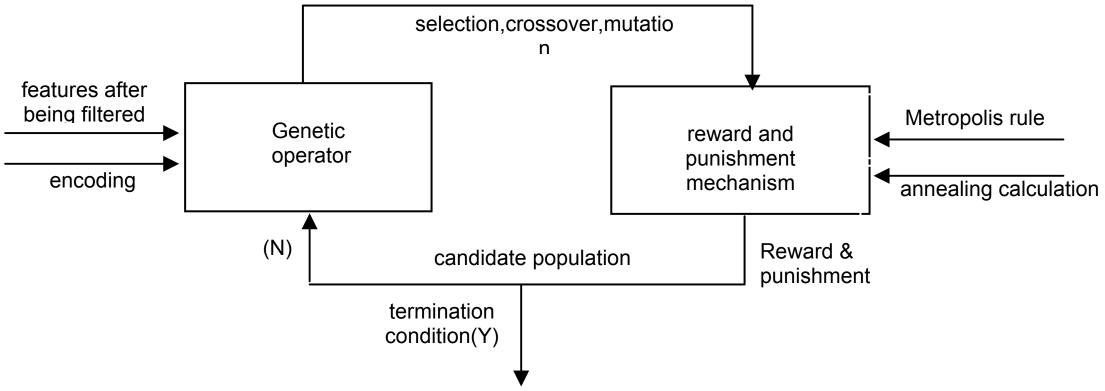

Figure 2. Improve GA Flow Diagram

In order to make up for the inadequacy of genetic algorithm's ability to mountain climbing, which often falls into local optimal solution but not the global optimal solution, we here take a reward and punishment mechanism as shown in Figure 2. We take the genetic algorithm as the main process and make simulated annealing algorithm integrated into it in order to is further adjust and optimize the groups. The core idea is: when the new individual fitness is higher than its parent fitness, we use the new to take the place of its parent as a reward for the new individual. Otherwise, as punishment, according to the Metropolis standard [8] we make the new to replace its parent by Boltzmann substitution probabilities P. According to the iterative improvement philosophy, it not only could enhance the global convergence but also accelerate the evolution speed and acquire the satisfied global optimal solution. The algorithm first determines an average value to reflect environment variable α in genetic algorithm operation.

$$\alpha = \frac{1}{k}\sum_{i=1}^{k}\left(f_i - \overline{f}\right)^2 \tag{3}$$

When α is larger, namely during the early period of genetic operator performing, individuals vary considerably and the phenotype variance α of genetic environment of is larger. The probability of acquiring the bad solution should be close to 1 and population update comes near to the full update mode. It could accelerate genetic algorithm convergence process and climb well over the local extremum obstacle to guide the search direction to get the global optimum solution. In the later period, α is smaller. We perform annealing operation on crossed and mutational individuals. It is not only beneficial to improve the capacity of searching global optimum but also to preserve excellent individuals. Population update is close to the covering update mode. It gradually turns into greedy search method until eventually finds the global optimal solution. So replacing probability is set as: exp(-θ⊿f/α) where θ>0. If α>=1 then θ=α+1 else θ=1.

3. Results and Analysis

Our experimental data is KDD Cup 1999 Data preprocessed by Columbia University. It provides the network connection data lasting 9 weeks acquired from a simulated local network. Each record in this set contains 41 dimension features and marks its category, such as Normal, Dos, U2R, Probing, R2L and so on. We extracts the data in the second and fourth week with the test platform being AMD3600+, 2G, XP operating system.

In order to make experiment operation convenient, we take samples from the data set and reduce the number of the instances. We respectively take random sampling method for DOS, PROBE, R2L, U2R, and NORMAL in training set to ensure the distribution consistency between samples and originals. Then we combine the 5 new samples to form the experimental training data set, including mixture of NORMAL and DOS, mixture of NORMAL and PROBE, mixture of NORMAL and R2L, mixture of NORMAL and U2R and mixture of DOS,PROBE,R2L and U2R. Each training data set has the same instance number of 11701. In these five training data set we use the feature selection algorithm proposed above and choose the corresponding feature subsets which respectively contain 5072, 5280 and 10158 records. In the test, we take 10-times running average as the test result, and then build the intrusion detection model based on all 41 features and selected feature subsets in every training data set. At last, we make comparisons in the aspect of system building time and test accuracy between 41 features based intrusion detecting system model and selected feature subsets based model.

The experiment first selects 33 features by mutual information indexes and then carries out the improved genetic algorithm. Let cross probability pc=0.6 , mutation probability pm=0.05 and iteration number is 100, according to the algorithm termination conditions the threshold value is 0.003. Experimental results show that we get the final algorithm optimal subset when the feature number is proposed finally in 18. This result shows that compared with original feature based system, the modeling time and detecting time both decline over half. It is shown in Figure 3.

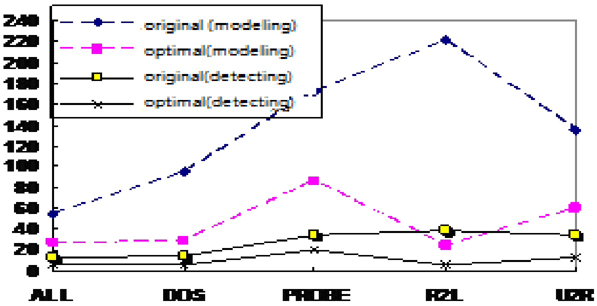

Figure 3. Comparison Time in the Modeling and Detecting between Original and Optimal Features

Compared with original features, the detection efficiency has greatly improved in the aspect detection time and false rate as shown in Table 1.

Table1. Detection Efficiency Comparison

	Original features	Optimal features
Feature number	41	18
Detection rate	98.504%	99. 129%
False rate	9.367%	7. 459%

In addition, we compare the improved genetic algorithm and traditional genetic algorithm in convergence speed and convergence precision: although there is no obvious increase in convergence time speed, we first use the mutual information indexes to screen the features and thus make the convergence accuracy more stable. It is proved to be superior to the traditional GA in the aspect of detection rate and false rate.

4. Conclusion

This paper utilizies a hybrid feature selection method to simplify intrusion detection features. It first adopts mutual information index and selects sequential forward construction method for eliminating redundant feature subset. Then on this basis, it uses improved genetic algorithm to construct the optimal subset. In the final simulation experiment, it also get the satisfied results of modeling time and detection rate.

References

[1] Zave P. *Classification of Research Effort s in Requirement's Engineering*. ACM Computing Surveys. 1997; 29(4): 315-321.

[2] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research.* 2004; 5(10):1205−1224.

[3] S Das. *Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection*. Proceedings of 18th Conference on Machine Learning. 2001; 74-81.

[4] E Xing, M Jordan, R Karp. *Feature selection for high-dimensional genomic microarray data*. Proceedings of Eighteenth International Conference on Machine Learning. 2001; 601–608.

[5] Lu Xin-guo, Lin Ya-ping, Chen Zhi-ping. An Improved Feature Selection Preprocessing Algorithm Based on Mutual Information. *Journal of Hunan University (Natural Science).* 2005; (2):104-107.

[6] Zhu Hong-ping, Gong Qing-ge, LEI Zhan-bo. Feature selection of intrusion detection based on genetic algorithm. *Application Research of Computers*. 2012. 29(4): 1417-1419.

[7] Ruggieri S. Efficient C4.5. *IEEE Transactions on Knowledge and Data Engineering*. 2002; 14(2): 438-444.

[8] Liu Yan, Han Cheng-de, Wang Yi-he, Li Xiao-ming. The Background of Simulated Annealing and The Monotonic Temperature Rising Sa. *Jouranl of Computer Research and development.* 1996; (01): 4-10.

[9] Wu Jian. Unsupervised intrusion detection based on feature selection. *Computer Engineering and Applications*. 2011; 47(26): 79-82.

[10] Zhang Yong, Cao Dong-xia. Novel improved fuzzy clustering algorithm applied in network intrusion detection. *Computer Engineering and Design*. 2012; 33(2): 479-483.

[11] Jing Xiao-pei, Wang Hou-xiang, NIE Kai, Luo Zhi-wei. Feature Selection Algorithm Based on IMGA and MKSVM to Intrusion Detection. *Computer Science.* 2012; 39(7): 97-99.

[12] Chen Wen, Zhao Yongjiu, Jun Zhouxiao, Compact and wide upper-stopband triple-mode broadband microstrip BPF. *Telkomnika*. 2012; 10(2): 353-358.

[13] Li Jun-Fanga, Zhang Bu-Han. Limitation of small-world network topology for application in non-dominated sorting differential evolution algorithm. *Telkomnika*. 2012; 10(2): 400-408.