Feature Selection Based on Rough Set and Modified Genetic Algorithm for Intrusion Detection[*]

Yuteng Guo,Beizhan Wang, Xinxing Zhao,Xiaobiao Xie, Lida Lin, Qingda Zhou
Software School of Xiamen University
Xiamen, China
Email: wangbz@xmu.edu.cn

***Abstract*—In the Network Intrusion Detection, the large number of features increases the time and space cost, besides the irrelative redundant characteristics make the detection accuracy dropped. In order to improve detection accuracy and efficiency, a new Feature Selection method based on Rough Sets and improved Genetic Algorithms is proposed for Network Intrusion Detection. Firstly, the features are filtered by virtue of the Rough Sets theory; then in the remaining feature subset, the Optimal subset will be found out through the Genetic Algorithm improved with Population Clustering approach for the best ultimate optimized results. Finally, the effectiveness of the algorithm is tested on the classical KDD CUP 99 data sets, using the SVM classifier for performance evaluation. The experiment shows that the new method improves the accuracy and efficiency in Network Intrusion Detection compared with the related researches of the intrusion detection system.**

***Key Words*—Intrusion Detection;Feature Selection; Rough Sets; Genetic Algorithm**

I. INTRODUCTION

The advent of Internet technology and its rapid development has brought much convenience to our lives, but also making people more and more relay on the Internet. More and more important information is spread through the network at this time, getting the information security becomes an important topic of Internet communication [1]. Various types of Intrusion Detection Systems (IDS) are used to detect network intrusions in real-time. For example: there is the network intrusion detection technology based on data mining (such as neural networks, support vector machine (SVM) [2], etc.). In the network intrusion detection, the system needs to handle massive amounts of network data in real-time manner, typically, the network data contains a large number of features[3], which significantly increases the load of IDS, but at the same time, there are many irrelevant and redundant features that will decline detection accuracy during the intrusion detection process based on machine learning mechanism and bring additional of the complexity of learning algorithms. All of these require IDS must be able to select the right subset of the most important features to improve the detection accuracy and efficiency.

For the Feature Selection in Intrusion Detection, the SNFS [2] algorithm used Neural network and Support Vector Machine. Later, CFSSGA [4] proposed a hybrid algorithm with correlation-based feature selection (CFS), and employed the SVM and genetic algorithm to achieve the optimization of intrusion detection. While, the FSRGA [1] algorithm is based on rough sets and improved genetic algorithms to improve feature selection. Both SNFS and CFSSGA algorithm need data classification for their each iteration. This produces much more time complexity and do not take care of the combination of characteristics as well as the balance of the number and classification accuracy. However, FSRGA algorithm did not optimize the genetic operation, which will easily to make the algorithm be trapped into a local optimal solution.

This paper introduces a new feature subset selection algorithm—FBRMGI (Feature Selection Based on Rough Set and Modified Genetic Algorithm for Intrusion Detection) based on rough set theory [5], genetic algorithms [6] and clustering [7], in fact, it combines the rough set and genetic algorithm based on clustering. Firstly, it uses the rough set theory to carry out the feature selection, then it uses the improved genetic algorithm based on clustering to find the optimal subset in the remaining feature subset. Finally, the genetic algorithm is meliorated with Population Clustering approach in pursuit for the more optimal result. In the end , the KDD CUP 99 [8] data set is used to test the effectiveness of the algorithm, and the SVM (Support Vector Machines) [9] classification is used to evaluate the effectiveness of the selected feature subset. The experimental results show that the new method improves the accuracy and efficiency in Network Intrusion Detection compared with the related researches of the intrusion detection system.

The rest of this paper is organized as follows: Section 2 describes the theoretical foundation; Section 3 introduces the FBRMGI algorithm; Section 4 describes the process of experiment and the analysis of experimental results; Section 5 introduces the Summary and points out the way for further research.

II. THEORETICAL FOUNDATION

A typical classifier-based intrusion detection process shown in Figure 1:

[*] This paper is sponsored by Fujian Provincial Youth Talents Program ”The Research of Complex-Intrusion-oriented Alert Information Aggregation and Association Analysis Technology(NO.2008F3101)” presided by A.P. Liang Shi

978-1-4244-6005-2/10/$26.00 ©2010 IEEE

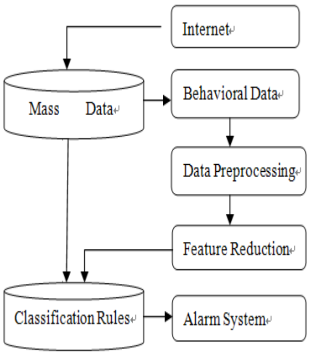


Figure 1. The process of intrusion detection

In the procedures above, the Mass data comes from the Internet, which is partly extracted as behavioral data, then the behavioral data is waiting for feature selection. Firstly, the extracted data is served with pretreatment, including data discretion and data formatting. After this, the data usually is of very high dimension. It is requisite to transform the data into a lower dimension feature space through feature selection in order to eliminate any irrelevant or redundant features for improving the detection accuracy and efficiency of IDS. The Feature Reduction module of Figure 1 plays that role. Then, classification rules are elicited according to the results of feature reduction. Finally, the Mass Data is classified complying with the classification rules. When the system detects any intrusion data which meets the alarm conditions, it will trigger the alarm system.

A. Feature selection

Feature selection will delete unimportant features according to certain rules in order to low the dimension of the feature space. Also, it can find the most effective subset of original feature set to improve prediction accuracy rate for classification and prediction models or to lower the complexity of the model structure in the guarantee of forecasting accuracy [10]. Feature selection methods can be divided into two categories in accordance with its independence to learning algorithms [11]: the feature selection methods independent to the learning algorithm are known as Filter method, otherwise known as the Wrapper methods.

In the Intrusion Detection feature selection study for the two classification problems (abnormal or normal): SNFH [2] algorithm takes advantage of the impact of every feature to the SVM classification accuracy to measure the importance of every feature, and takes it as a basis for classification of features. While CFSSGA [4] feature selection algorithm is a mixture of CFS, SVM, and GA. firstly, it generates sub-cluster with genetic algorithm, using CFS to filter out the best feature subset, which is then evaluated by the SVM classifier. At the same time, FSRGA [1] algorithm computes the importance of each feature through rough set, which is then introduced into initial population operator of the genetic algorithm as heuristic information, while the number and Classification capability of the features are incorporated into fitness function of genetic algorithm.

SNFH and CFSSGA have adopted Wrapper feature selection method, but the massive computing reduces the efficiency of the algorithm, and they do not take care of the combination of features or the balance of the number of features and classification accuracy. While FSRGA adopts Filter feature selection method to improve the efficiency of the algorithm but without optimization to the genetic operation, making the algorithm easy to be trapped into a local optimal solution. Based on the above analysis, FBRMGI algorithm is proposed in this paper with filter method.

B. Rough set theory and information theory

Rough set theory [5] is proposed by professor Pawlak at Warsaw University of Information Technology and Management to deal with incomplete information. It does not require any priori information, and can effectively analyze and deal with incomplete, inconsistent, inaccurate data. It defines the knowledge from a new perspective: taking the knowledge as the partition of universe, where the equivalence relations are used to formally express the classification. Through large amounts of data analysis, this method deletes some relative information in accordance with the relationship of the two equivalence relations in the universe and extracts potential valuable knowledge of the rules. This method has been widely used in knowledge acquisition, rule extraction, machine learning, decision analysis, pattern recognition, data mining and other fields [10], it is very suitable for the safety rules learning and detection.

Information theory is a branch of application mathematics and electrical engineering involving the quantification of information. It was developed by Claude E. Shannon to find fundamental limits on compressing and reliable storing and communicating data.

Suppose U is a universe, P, Q are the equivalence relation in it (i.e.: knowledge), according the information theory, the entropy of knowledge P, is defined as [12]:

$$H(P) = -\sum_{i=1}^{n} p(X_i)\log p(X_i) \tag{1}$$

The conditional entropy of knowledge Q relative to P, is defined as [12]:

$$H(Q \mid P) = -\sum_{i=1}^{n} p(X_i)\sum_{j=1}^{m} p(Y_j \mid X_i)\log p(Y_j \mid X_i) \tag{2}$$

Among them: $p(Y \mid X) = card(Y \cap X) / card(X)$.

Mutual Information is defined as [12]:

$$I(R,D) = H(D) - H(D \mid R) \tag{3}$$

According to the definition of mutual information, feature important degree is defined as [12]:

$$\begin{aligned} SGF\ (a,R,D) &= I(R \cup \{a\};D) - I(R;D) \\ &= H(D \mid R) - H(D \mid R \cup \{a\}) \end{aligned} \tag{4}$$

C. *Genetic Algorithm*

The genetic algorithm (GA) is a search technique used in computing to find exact or approximate solutions to optimization and search problems. Genetic algorithms are categorized as global search heuristics. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology [6]. The main steps of GA are as follows:

Initialization: Generate Initial population.

Selection: Calculate individual's fitness value based on fitness function and retain some candidate solutions as well as give up other candidate solutions According to fitness.

Reproduction: Mutation and Crossover.

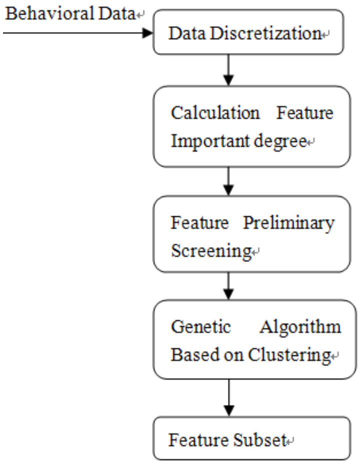
Termination: Before the termination conditions are met, repeat the processes of selection, Mutation and crossover.

D. *Clustering*

The process of putting Physical or abstract objects into a collection of similar object is called clustering [7], but different from the classing, label of each object is unknown. Clustering puts the data objects into categories or clusters, making the objects in a cluster of much similarity, while objects in different clusters in a high degree of dissimilarity. Usually, metric distance is used to measure the dissimilarity.

III. FEATURE SELECTION ALGORITHM --FBRMGI

FBRMGI algorithm flow is shown in Figure 2:



Behavioral Data
Data Discretization
Calculation Feature Important degree
Feature Preliminary Screening
Genetic Algorithm Based on Clustering
Feature Subset

Figure 2. FBRMGI algorithm flow

As shown in Figure 2, this algorithm includes Four Modules: Data Discretization, Calculation Feature important degree, Feature Preliminary Screening and Genetic Algorithm Based on Clustering. The following is detailed description of these modules:

A. *Data Discretization:*

Rough set requires that the data must be Discretization. In this paper, the Naïve Scaler algorithm [13] is used to make the continuous features become disserted features.

B. *Calculation Feature Important degree:*

Using formula (4), calculate the Mutual Information $SGF(a,\varnothing,D)$ of each feature a against the decision-making feature.

C. *Feature Preliminary Screening:*

Firstly, for the behavioral data, statistical methods are used to get the classification of each feature, then the features which just have one value will be find. Those kinds of features will not impact the decision-making feature, so they can be deleted.

Then, according to the importance of each feature, the features whose feature important degree are below a certain minimal threshold value will be deleted, while retaining the features which have a higher important degree. Here it supposes the number of features which are reserved is M .

D. *Genetic Algorithm Based on Clustering*

1) Population initialization:

Initial population uses randomly selected method. Population coding accesses binary coding scheme, 1 refers that the feature was selected while 0 is not. The length of chromosomes N equals to the number of features while the features which just have one value, the features with low importance and the features with high importance is removed.

2) The definition of fitness function :

The fitness function depends on three aspects:

a) The number of features in feature subset N , the less the higher fitness;

b) The number of features which have high feature importance-M, the larger the higher fitness;

c) The mutual information of feature subset and decision-making features, the more the higher fitness.

Therefore, the fitness function of chromosomes can be defined as:

$$f(x) = \frac{N - Lx}{N} + \frac{Mx}{N} + IRD(x) \tag{5}$$

Where, N refers to length of chromosomes. Lx refers to the number of 1 in the chromosomes. Mx refers to the number of features which have high feature importance. $IRD(x)$ refers to the mutual information of the chromosome x against decision-making feature.

3) Selection :

It uses the strategies of roulette and elite retention. For parent populations individuals, some of the best individuals are chosen into the next generation, while the remaining individuals using roulette selection method according to the ratio of their fitness in the sum of the whole population of individual fitness.

4) Population clustering :

According to biological evolution, the crossover rate and mutation rate should be modest and dependent on the evolution of groups [7]. In order to prevent the local premature optimization of population and improve the population performance, populations clustering is used to building self-adaptive crossover and mutation rate. Firstly, use the clustering method to cluster the population. The similar Chromosomes are classified in a class with the corresponding mutation rate, and the crossover rate is shared by all classes; secondly, according to the number of populations and each category's fitness the values of the crossover rate and mutation rate are determined.

Clustering approach uses the K-MEANS algorithm [7], the steps are as follows:

a) Randomly selecte k objects, each object represents center or initial value of the cluster;

b) Calculate the cluster centers of the k classes.

c) Assign the samples to the most recent cluster center, re-calculate the center of each class;

d) Repeat the previous two steps, until members of the class which belongs with no change in a certain range.

Here, similarity of the sample to the cluster center is calculated using Manhattan distance. The formula for calculation is as follows:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{in} - x_{jn}| \quad (6)$$

Each individual is an n-dimensional Vector. $d(i,j)$ represents the distance between the 2 n-dimensional vectors. Recalculate the cluster centers using the following strategies: find the mode of each bit of all individuals in each cluster. And form the new cluster center using these modes. Mutation rate and crossover rate adjusting tactics are based on the following rules:

If $f > f_a$ And $M > M_a$ Then $P_m = P_{m\max}$ $P_c = P_{c\min}$

If $f > f_a$ And $M < M_a$ Then $P_m = P_{m\min}$ $P_c = P_{c\min}$

If $f < f_a$ And $M < M_a$ Then $P_m = P_{m\min}$ $P_c = P_{c\max}$

If $f < f_a$ And $M > M_a$ Then $P_m = P_{m\max}$ $P_c = P_{c\max}$

M represents number of individuals of a cluster. M_a represents the average number of individuals in each cluster. f represents the fitness value of individual which is the center of a cluster. f_a represents the average fitness of entire population, P_m represents the mutation rate of one cluster, P_c represents the crossover rate of the entire population, $P_{m\min}$ 、 $P_{m\max}$ represents the minimum and maximum of mutation rate, $P_{c\min}$ 、 $P_{c\max}$ represents the minimum and maximum of crossover rate.

5) Crossover 、 Mutation and Termination

According to biological evolution, a distant relative breeding can enhance the diversity of population. According to the clustering results of the previous step, selecting the individuals in different clusters and using single-point crossover strategy for crossover based on the current cross-rates are very well. Then the genes bit involved in variation is selected randomly and inverted according to the mutation rate in the cluster that is, mutating the individual. Until the maximum fitness of population individual is prone to be convergent after certain generation, the iteration is terminated.

IV. EXPERIMENTS

KDD CUP 1999 [8] is used as the experimental data, and SVM classifier is employed on the selected feature subset in this paper and literature [1] [2] [4]. The efficiency of feature subset results are compared.

A. Experimental data

KDD CUP 1999 data set [8] is from MIT, created by LINCOLN laboratory for constructing the connection record and extract features on the foundation of 1998 DARPA intrusion detection data set. The data extracts 41 features, including 34 continuous features and 7 discrete features [8].

In order to accurately evaluate the FBRMGI feature selection algorithm, this paper uses the corrected data set from KDD CUP 1999 as the experimental data sets. Data sets are divided into normal and abnormal type. All the ones of the different attacks behavior are marked as abnormal. 20000 records are collected for feature selection, anther 80000 records are collected for the evaluation of feature subset including 50000 records for training set of SVM, the rest 30000 records for testing set of SVM.

On the basis of the KDD CUP99 data sets, SNFS [2] chooses 13 features as feature subset (in feature ID indicated): <1, 2, 3, 5, 6, 9, 23, 24, 29, 32, 33, 34, 36>; RMGFS [1] chooses 12 features as feature subset: <1, 6, 12, 14, 23, 24, 25, 31, 32, 37, 40, 41>; CFSSGA [4] chose 12 features as feature subset: <2, 3, 5, 6, 7, 12, 15, 28, 30, 34, 37, 40>. these three results are compared.

B. Experimental configuration

Experimental configuration is as follows: the minimum threshold value for feature importance is 0.0001; the size of initial population is 100; the Maximum for mutation rate is 0.75; the Minimum for mutation rate is 0.08; the Maximum for crossover rate is 0.67; the Minimum for crossover rate is 0.2.

Experimental environment is as follows: Intel(R) Pentium(R) M; CPU 1.86GHz; 1.50GB Memory; Windows XP System.

C. Feature selection results

FBRMGI feature selection algorithm for feature subset selection is shown in TABLE Ⅰ. <1,2,3,5,6,23,33> are the same compared with the SNFS [2] algorithm's feature subset. <1,6,23,31> are the same compared with the CFSSGA [4] algorithm's feature subset. <2,5,6> are the same compared with the RMGFS [1] algorithm's feature subset.

5 groups of experiments are carried out on original features, feature subset of SNFS [2] , feature subset of CFSSGA [4], feature subset of RMGFS [1] and FBRMGI. There are five indicators used to evaluate the results: the number of features, Detection time, Detection rate, Accuracy rate, False rate.

$$DR = \frac{TE}{TE_m} \tag{7}$$

$$\mathrm{A}R = \frac{T\mathrm{A}}{T\mathrm{A}_m} \tag{8}$$

$$Mis = \frac{F}{T_m} \tag{9}$$

TABLE I. RESULTS OF FEATURE SELECTION

ID	Name	Feature ID	Description
1	duration	1	length (number of seconds) of the connection
2	protocol type	2	length (number of seconds) of the connection
3	flag	4	normal or error status of the connection
4	src_bytes	5	number of data bytes from source to destination
5	dst_bytes	6	number of data bytes from destination to source number of
6	num_failed_logins	11	failed login attempts
7	is_guest_login	22	1 if the login is a "guest"login; 0 otherwise
8	count	23	number of connections to the same host as the current connection in the past two seconds
9	srv_diff_host_rate	31	% of connections to different hosts
10	dst_host_srv_count	33	number of connections to the same service as the current connection in the past two seconds
11	dst_host_diff_srv_rate	35	% of connections to the different service

DRrefers to Detection rate; TE refers to the number of samples which are labeled with abnormal in detection, TE_m refers to the number of samples which are abnormal; $\mathrm{A}R$ refers to Accuracy rate, $T\mathrm{A}$ refers to the number of samples which are classified correctly, $T\mathrm{A}_m$ refers to the total number of samples in the detection; Mis refers to False rate, Frefers to the number of samples which are normal but labeled with abnormal in detection, T_m refers to the number of normal samples. The results are shown in TABLE Ⅱ.

As show in TABLE Ⅱ, the optimized feature subset has better performance. Detection rate has increased from 95.2711% to 98.2133%, Accuracy rate has increased from 97.66% to 98.2133%, False rate has reduced to 0.9191% from 7.6322%.The number of features in FBRMGI is less than in SNFS [2] , CFSSGA [4] and RMGFS[1].The Detection rate, Accuracy rate and False rate of FBRMGI are better than the others.

TABLE II. RESULTS OF DETECTION

	Original features	SNFS [2]	CFSSGA[4]	RMGFS[1]	FBRMGI
Number of features	41	13	12	12	11
Detection time	105s	57s	49s	58s	54s
Detection rate	95.2711%	97.2566%	96.9499%	97.7466%	98.409%
Accuracy rate	98.2033%	98.19%	97.66%	97.92%	98.2133%
False rate	7.6322%	6.4277%	4.9584%	4.4224%	0.92%

V. CONCLUSIONS

Based on rough set theory [5], genetic algorithms [6] and clustering [7], a new feature selection algorithm FBRMGI combining of rough set and genetic algorithm on foundation of clustering is proposed. Firstly, use the rough set theory to process feature selection, then use the improved genetic algorithm based on clustering to find the optimal subset in the remaining subset, in the end, combine the results of the first two steps to get the final results. The Results of experiments show that FBRMGI is better at Detection rate, Accuracy rate, False rate than the other three algorithms, meanwhile, the number of feature is less. Building more rigorous mathematical formulas for crossover rate and mutation rate to design a more reasonable experimental parameters will be our work forward.

ACKNOWLEDGEMENT

This paper is sponsored by Fujian Provincial Youth Talents Program "The Research of Complex-Intrusion-oriented Alert Information Aggregation and Association Analysis Technology(NO.2008F3101)" presided by A.P. Liang Shi

REFERENCES

[1] Luyin Chen,Qingshan Jiang,Lifei Chen. “A Feature Selection Method for Network Intrusion Detection”.computer Research and Development Supplement,45(10):156-160,2008

[2] A.H.Sung, S.Mukkamala, “Identifying Important Features for Intrusion Detection using Vector Machines and Neural Networks”, Proceedings of International Symposium on applications and the Internet Technology, pp. 209-216, 2003

[3] L.Chen, L.Shi, Q.Jiang and S.Wang. “Supervised Feature Selection for Dos Detection Problems Using a New Clustering Criterion”, Journal of Computational Information Systems, 3(5):1983-1992

[4] Shazzad, K.M., Jong Sou Park, “Optimization of Intrusion Detection through Fast HybridFeature Selection”, Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005. Sixth International Conference on , vol., no., pp. 264-267, 05-08 Dec. 2005

[5] Pawlak.Z, “Rough Sets and Intelligent data analysis”,Computer and Information Science,2002

[6] Holland.J.H, “Adaptation in Natural and Artificial Systems”.University of Michigan Press,Ann Arbor,1975

[7] Han Jiawei, Micheline Kamber. “Data mining: concepts and techniques”. 2nd ED. Beijing :China Machine Press, 2006

[8] KDD CUP 1999 DataSet. http: // kdd. ics. uci. Edu / databases /kddcup99/task.htm

[9] C.J.C.Burges. “A tutorial on support vector machines for pattern recognition”. Data Mining and knowledge Discovery, 2(2):121-167,1998

[10] D.Koller, M.Sahami. “Toward optimal feature selection”, Proceedings of the rnational Conference on Machine Learning , 1996

[11] K.M.Shazzad, J.S.Park.“Optimization of Intrusion Detection through Fast Hybrid Feature Selection”, Proceedings of the Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT’05),2005.

[12] Yanhuai Ma. “Data Mining Methods based on Rough Set Theory ". PhD thesis, Chinese Academy, 2003

[13] ROSETTA SOFTWARE. Http://tosetta.kb.uu.se/general