Intrusion Classifier based on Multiple Attribute Selection Algorithms

Weiwu Ren
Jilin University, Changchun, 130012, China
Email: Renww339@163.com

Liang Hu
Jilin University, Changchun, 130012, China
Email: hul@jlu.edu.cn

Kuo Zhao
Jilin University, Changchun, 130012, China
Email: zhaok@jlu.edu.cn

Jianfeng Chu
Jilin University, Changchun, 130012, China
Email: chujf@jlu.edu.cn

Bing Jia
Jilin University, Changchun, 130012, China
Email:Bing_jia@qq.com

***Abstract*— with the rapid growth of attack patterns, the number of attributes for detecting attacks gradually increased. Moreover, an automatic attack classification method, as the next thing of intrusion detection, is needed. For solving the above problems, an intrusion classifier based on multiple attribute selection algorithms has been proposed. The classifier includes various combinations with different representative attributes selection algorithms and classification algorithms. A series of experimental results on well-known KDD Cup 1999 data sets indicate real time performance and classification performances of different combinations.**

***Index Terms*—attribute selection, attack classification, attack patterns**

I. INTRODUCTION

With the rapid growth of attack technology, attack pattern has become from simple to complicate. For various attack patterns, it is impossible to classify them only by use of security administrator's experience and manual classification. Moreover, the number of attribute which is used to classify attack patterns increases gradually, which leads to low real time and low classification performance.

Many representative intrusion detection algorithms [1-3, 11-13] based on machine learning have been proposed. Generally, in order to maintain high real time and low false positive, most of algorithms are based on two categories, namely, normal and anomaly. Many unsupervised intrusion detection algorithm based on data mining [4-5] can handle them well. However, how to label further attacks or how to classify them is another urgent problem for increasingly complex security system. The traditional classification uses the well-known or obvious attributes to match and identify the known attack pattern. The main advantage of this method is that it can accurately and efficiently classify all known attacks. But it totally depends on the security administrator's rich classification experience. And this method has poor efficiency. For this, many classification algorithms based on data mining are introduced for classifying attack patterns. Attack classification, as the next work of intrusion detection, has been valued gradually.

With the rapid growth of attack patterns, the number of attribute increased. The intrusion classification algorithm based on data mining is sensitive to high dimension of attributes. For solving this problem, many attribute selection algorithms [6-8] have been proposed. According to different evaluation methods, attribute selection can be divided into two classes: filter [9] and wrapper [10]. In the filter, the attribute itself, as the evaluation criteria is calculated by different evaluation algorithms. Compared with filter, the wrapper adopts the classification error of machine learning algorithms as the evaluation criteria. Generally, the speed of filter is faster and its selection result has nothing to do with machine learning algorithm. But its result is less effective. Wrapper has a relatively slow speed and needs a lot of computation. Its result totally depends on its classification algorithm. Its selection result has good performance.

doi:10.4304/jcp.8.10.2536-2543

For purposes of reducing dimensions of data and testing performance of different classification algorithm in different dimensions, an intrusion classifier based on multiple attribute selection algorithms has been proposed. The system includes two parts: attribute selection and classifier. In attribute selection part, there are four representative algorithms: Information Gain Rate (IGR), Correlation Feature Selection(CFS), Support Vector Machine(SVM) and decision tree. In the classifier part, there are two main classification algorithms: SVM and C4.5. Different attribute selection algorithm and different classification algorithms are combined.

The rest of paper is organized as follows. In section 2, we introduce the framework of system. Section 3 presents details of attribute selection. Section 4 presents details of classifier. Section 5 presents our experiments results and analysis. Finally, we summarize our conclusions and future work in section 6.

II. FRAMWORK OF SYSTEM

As is shown in the figure 1, the system is made of two parts: attribute selection part and intrusion classifier part. The feature part has for sub modules: IGR, CFS, SVM and C4.5. The intrusion classifier part has two sub modules: C4.5 and SVM. Different attribute selection modules can combine with intrusion classifier module. So there are 8 different combinations. Because attribute subset evaluation algorithm totally depends on its classification algorithm, the SVM-C4.5 and C4.5-SVM has no practical meaning. In this paper, we mainly study 6 combinations of them. They are respectively IGR-SVM, CFS-SVM, IGR-C4.5, CFS-C4.5, SVM-SVM and C4.5-C4.5.

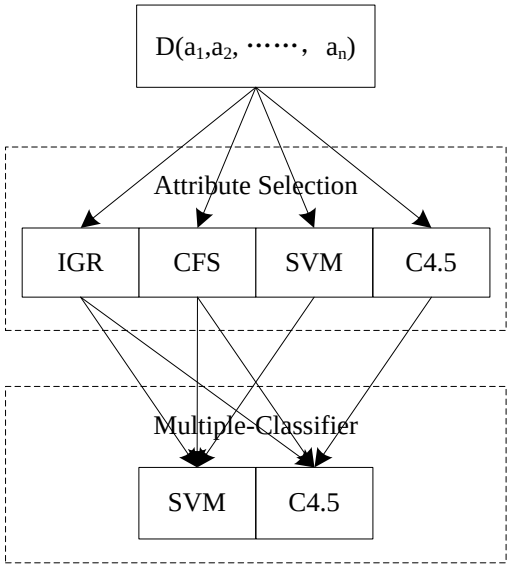

Figure 1. Framework of System

III. ATTRIBUTE SELECTION

According to different evaluation methods, attribute selection can be divided into two classes: filter and wrapper. In the filter, the attribute itself, as the evaluation criteria is calculated by different evaluation algorithms. Compared with filter, the wrapper adopts the classification error of machine learning algorithms as the evaluation criteria. Generally, the speed of filter is faster and its selection result has nothing to do with machine learning algorithm. But its result is less effective. Wrapper has a relatively slow speed and needs a lot of computation. Its result totally depends on its classification algorithm. Its selection result has good performance.

According to different evaluation targets, attribute selection can be divided into two classes: single attribute evaluation and subset evaluation. Single attribute evaluation evaluates every attribute. Evaluation result is that every attribute has its evaluation value. And all attributes are ranked by evaluation values. The evaluation value can be used to weight or to select attributes by setting a threshold. Attributes in the low order positions may be eliminated. Subset evaluation searches the best subset by traversing the whole attribute space. The evaluation quality is measured by the attribute subset evaluator. The attribute subset evaluator can be filter or wrapper.

A. Attributes Evaluation

The basic work flow of single evaluation is shown as in the figure 2. It can be divided into three steps:

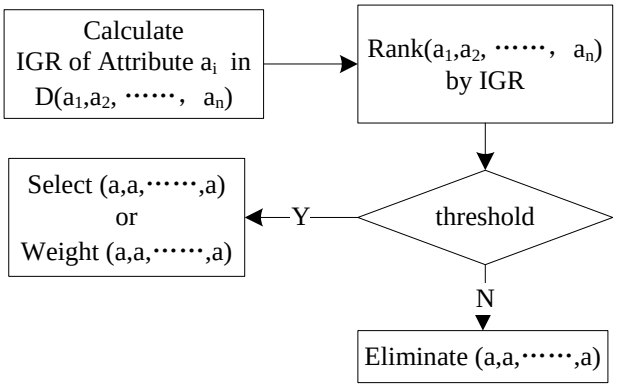

Figure 2. Workflow of Single Evaluation

(1) Calculate evaluation criteria (EC): EC depends on impurity measure of every attribute. Impurity measures include information gain, information gain ratio, Gini-index, distance measure, J-measure, G-statistics, χ^2 -statistics, weight of evidence, MLP, orthogonality measure, relevance and relief.

(2) Rank attributes: According to evaluation value of every attribute, all attributes are ranked.

(3) Select or weight attributes: A threshold can be set based on actual needs. If some attributes in the low order positions are less than the threshold, attributes would be eliminated and the aim of attribute selection is achieved. Or attributes can be weighted by the evaluation value.

The evaluation criterion is information gain ratio. Information gain ratio is defined as: the set S is split into

the n sub set $\{S_1,S_2,\cdots\cdots, S_n\}$ by the discrete attribute A. Information gain ratio is shown as follows:

$$GainRatio(S,A)=\frac{Gain(S,A)}{SplitI(S,A)} \quad (1)$$

$$fit(\alpha_i)=CR(\alpha_i)+CR_{\max}-2*CR_{avg}\ /\ |S| \quad (CR_{\max}+CR_{\min}>2*CR_{avg}) \quad (4)$$

Here, α_i is the individual of ith population. $CR(\alpha_i)$ is the classification rate of subset. S is all subsets. |S| is number of S. $CR_{\max}$ is the maximum of classification rate of the i[th] generation. $CR_{\min}$ is the minimum of classification rate of the i[th] generation. CR_{avg} is the average classification rate of the i[th] generation.

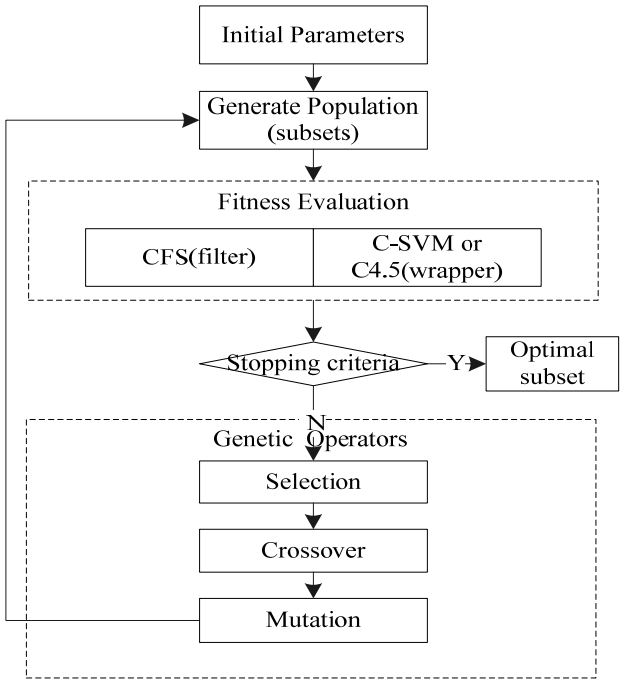

Figure 3. Workflow of Subset Evaluation

Information split splitI(S,A) is defined as:

$$SplitI(S,A)=-\sum_{i\in Value(A)}\frac{S_i}{S}\log_2(\frac{S_i}{S}) \quad (2)$$

B. Subset Evaluation

The basic work flow of subset evaluation is shown as in the figure 3. It can be divided into four steps:

(1) Encode parameters

Parameter coding is the key of implementing genetic algorithm. In this paper, attributes of subset, as the only parameter, are encoded.

(2) Generate subsets

According to initial parameters, the initial population is randomly generated and distributes homogeneously in solution space.

(3) Fitness function

Genetic algorithm aims to search the optimal fitness function, but attribute selection aim to achieve the optimal minimum of various performance indicators. So the fitness function is shown as follows:

$$fit(\alpha_i)=CR(\alpha_i)-CR_{\min}\ /\ |S| \quad (CR_{\max}+CR_{\min}>2*CR_{avg}) \quad (3)$$

Or

(4) Genetic operators

Genetic operator is the key of searching the optimal. There are three classes of genetic operator: selection, crossover and mutation.

After crossover, fitness of individual is accumulated. When it gets to the threshold, the individual which is accumulated can generate population, but the other are eliminated.

A pair of chromosomes is mated and the position of crossover is randomly generated. Then the chromosome breaks in the position of crossover. One chromosome changes into two parts: chromosome 1 and chromosome 2. The other also changes into two parts: chromosome 3 and chromosome 4. The chromosome 1 and chromosome 4 are mated and the new chromosome is generated. So is chromosome 2 and chromosome 3.

One gene of chromosome is mutated in a smaller probability and new chromosome is generated.

There are two conditions of algorithm convergence: (1) optimal subset tends to be stable (2) the generation number has exceeded the threshold.

There are two kinds of evaluation criteria: filter and wrapper. Correlation Feature Selection (CFS) is one kind of filter. SVM and C4.5 belong to wrapper.

The Correlation Feature Selection (CFS) measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other". In order to prove this theory, the following definition and theorem are introduced:

The definition 1: R_n is a N-dimensional features space; X_n is a mode vector of R_n; $F^n = \{x_j^n\}_{j=1}^N$ is the total number of samples of m pattern recognitions, and $\sum_{i=1}^{m} N_i = N$, N_i is the number of samples of the i^{th} mode. The Fisher ratio criterion is defined as $J_n(\partial^n) = (\frac{\partial^{n^T} S_b^n \partial^n}{\partial^{n^T} S_i^n \partial^n})$.

In definition 1, S_b^n and S_i^n is respectively scatter matrix between classes and scatter matrix in class. They are nonnegative definite matrix. ∂^n is the optimal classification identification vector which makes equation (1) be maximum value. Equation (1) represents the classification capability of ∂^n.

Theorem 1: R_{n+1} is a new feature space which consists of the old R_n and the other related feature. The column vector d_1 of H^{n+1} and the column vector $Y_1^N, Y_2^N, \cdots Y_n^N$ of R_n are linear correlation, then Theorem 1 indicates that if new linear correlation feature is inserted into the old feature space, the new feature space has the same classification identification capability with the old one. That is also said, if the linear correlation feature is deleted from the old feature space, the obtained space has the same classification identification capability with the old one. The following equation gives the merit of a feature subset S consisting of k features:

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (5)$$

Here, s is a subset which contains k features, $Merit_s$ is an evaluation result of correlation of S. $\overline{r_{cf}}$ is the average value of all feature-classification correlations, and $\overline{r_{ff}}$ is the average value of all feature-feature correlations. When $\overline{r_{cf}}$ is high and $\overline{r_{ff}}$ is low, the classification effect is good. The CFS criterion is defined as follows:

$$CFS = \max_{S_k}\left[\frac{r_{cf1} + r_{cf2} + \cdots r_{cfk}}{\sqrt{k + 2(r_{f_1f_2} + \cdots + r_{f_if_j} + \cdots + r_{f_{k1}f_1})}}\right] \quad (6)$$

SVM and C4.5 belong to wrapper and are also two kinds of classifier, which would be introduced in the section 4.

VI. Analysis of Evolution Mechanism

A. Multiple-classifier based on SVM

As a self-organized complex system, telecom industry system has nonlinear characteristic. In the process of evolution, telecom industry system introduces the factors of policy, technology, capital and demand from the environment as negative entropies. Through the amplification of the nonlinear interaction in the inside and outside of the system, a chain reaction of the telecom system and even the socio-economic system is caused. The telecom industry is driven further from equilibrium and into a nonlinear zone. The growth rate of network value begins to have nonlinear variations, showing a nonlinear increase in the number of users. Ultimately, a tremendous influence and impact on the evolution of the telecom industry is caused.

Support vector machine is based on the linear division. But in order to solve nonlinear division, all data of nonlinear division in the raw space would be mapped into new data in the high-dimensional spaces, which make new data in the high-dimension spaces be linear division.

In this paper, nonlinear soft margin support vector machine (C-SVM) is the one of classification algorithm. And it can be also the evaluation function in the wrapper mode.

Although the training data can be mapped into high-dimension spaces by kernel function in nonlinear hard margin support vector, few kernel functions can ensure them to be linear division in any case. Therefore, it is reasonable for introducing slack variable of nonlinear soft margin support vector. The original problem is transformed into: $\tilde{T} = \{(\tilde{x}_1, y_1), \cdots (\tilde{x}_l, y_l)\}$

Here: $\tilde{x}_i = \phi(x_i)$, $(\tilde{w} \cdot \tilde{x}) + \tilde{b} = 0$

$$\min_{w,b} \frac{1}{2}\|\tilde{w}\|^2 + C\sum_{i=1}^{l} \xi_i$$

$$s.t. \quad y_i((\tilde{w} \cdot \tilde{x}_i) + \tilde{b}) + 1) \geq 1 - \xi_i, i = 1, \cdots, l$$

Dual problem:

$$\begin{cases} \min\limits_{\alpha} \quad \frac{1}{2}\sum\limits_{i=1}^{l}\sum\limits_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) - \sum\limits_{j=1}^{l} \alpha_j, \\ s.t. \quad \sum\limits_{i=1}^{l} y_i \alpha_i = 0 \qquad 0 \leq \alpha_i \leq C \end{cases}$$

$$b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x_j)$$

$$f(x) = \mathrm{sgn}\left(\sum \alpha_i^* y_i K(x, x_i) + b^*\right)$$

TABLE I.
OPTIMAL SUBSETS

Algorithm	Optimal Subsets
IGR	3,15,19,18,13,5,4,25,17,23,28,34,31,2 (merit>0.4)
CFS	3,5,13,15,23,24,25,26,33
J48	2,3,4,7,10,13,16,22,24,25,27,29,32,35,36,38,40,41
SVM	1,2,3,4,5,7,11,12,13,15,16,17,19,20,21,24,25,29,30,31,32,33,34,38,39

SVM is a dichotomous classifier. To solve the M classes-classification, M(M-1)/2 SVM classifiers need to be built. Every classifier votes and the class which has most votes is the final class.

B. Multiple-classifier based on Decision Tree

Compare with ID3, C4.5 selects the most information gain attribute as measuring attribute. And C4.5 uses the information gain ratio as measuring standard. Information gain ratio is the ratio of information gain to information split. ID3 only can deal with discrete data. C4.5 split the continuous values into discrete interval. So C4.5 not only deals with discrete data, but also continuous data.

One decision tree should be built and refined by training. The process of building decision tree is the process of machine learning. The process of building tree is shown as follows:

(1) Create one node N, the tree starts with this node.

(2) If all the instances belong to the same class, this node is leaf node. And this node is marked by the class number.

(3) For every attribute, if its data is continuous, the data should be dispersed.

(4) The information gain ratio of every attribute should be calculated. The highest information gain ratio should be selected and marked.

(5) The consistent value of every branch attribute should be calculated. A branch which has the same value is generated.

(6) S is the set of branches of training data set. If S is null, a leaf node should be added and marked as the class.

(7) If S is not null, the above steps are recursively called.

Usually the decision tree needs to be pruned. It can help reduce the size of tree and improve the classification accuracy. But in the application of intrusion detection, it is impossible to avoid the fallible data by pruning branches. The process of pruning branches is shown as follows:

(1) Calculate classification error. The classification error is defined as the sum of weights of data points which are not belong to this node. For the Non-leaf node, the classification error is the sum of classification errors of its child nodes.

(2) Decision. if the classification error is more than the attribute which appears most in this node, the branches of this node is pruned and this node become the leaf node. And the most attribute value is assigned to this node.

V. ANALYSIS OF EVOLUTION MECHANISM

A. Dataset

KDD CUP 1999 data set which is deprived from 1998 DARPA Intrusion Detection Evaluation program held by MIT Lincoln Labs, is employed to study the utilization of machine learning for intrusion detection by numerous researchers. The dataset includes all kinds of simulated intrusion actions in the complicated network environment, where each connection instance contains 41 features. In this paper the KDD CUP 1999 data set have been selected as the simulated traffic source of our experiments. 100,000 connection instances, as the

TABLE II.
TIME PERFORMANCE

Algorithms	Time performance	
	Training time	Testing time
ALL-SVM	147.8s	48.8s
CFS-SVM	48.3s	18.3s
IGR-SVM	69.2s	25.7s
SVM-SVM	92.6s	36.3s
ALL-J48	22.9s	10.2s
CFS-J48	17.4s	7.3s
IGR-J48	20.3s	8.9s
J48-J48	19.8s	7.8s

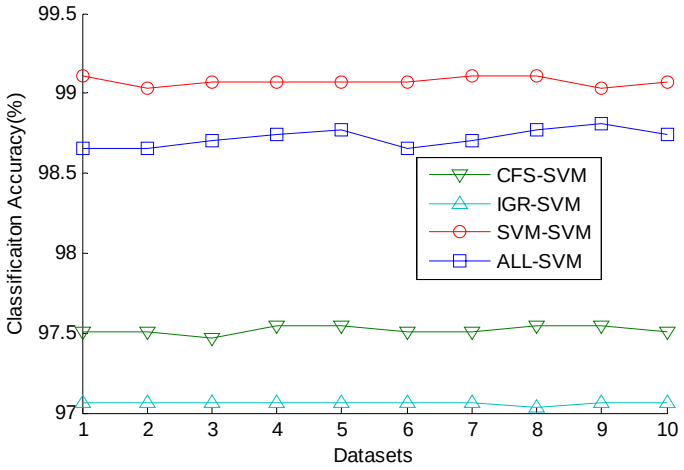

Figure 4. Classification Accuracies of SVMs

training dataset, are extracted randomly from the file kddcup. data_10_percent. 100,000 connection instances, as the predicting dataset, are extracted randomly from the file corrected. All data in the training dataset and testing dataset would be classified and labeled four classes: Probe, DoS, U2R and R2L.

B. Results of Attribute Selection

In table 1, optimal subsets which are selected by four algorithms are shown. Here optimal subset of CFS has the least number of attributes. Optimal subset of SVM has the most number of attributes. The optimal subset of IGR is obtained by the criteria: merit>0.4. It can be concluded from table 1 that the 3^{th} attribute is most important in the KDD CUP 99. The 3^{th} attribute appears in all optimal subsets. The aim of selection attribute is to find out the most important attribute and the optimal subset.

C. Comparison of Real Time

In table 2, there are 6 different attribute selection algorithm and 2 full attribute algorithm. ALL-SVM and ALL-J48 represents two full attribute algorithms. ALL- represents the full attributes or no attribute selection in the front. SVM and J48 represent two basic classifiers. In a similar way, CFS-SVM represents that attribute selection is CFS in front and the classifier is SVM in back. It can be concluded from table 2 that after attribute selection, training time and testing time reduces. Especially CFS has the minimal subset, two classifiers based on CFS: CFS-SVM and CFS-J48 has the optimal real time performance. The experiment result of table 2 indicates that the number of attributes directly influences on the real time performance.

D. Comparison of Classification Performance

As shown in the figure 4, SVM-SVM has the best classification performance and IGR-SVM has the worst classification performance. After selecting attributes by CFS and IGR, their classification accuracies are lower than ALL-SVM. By contrast, SVM-SVM has less attributes and higher classification accuracy. The most

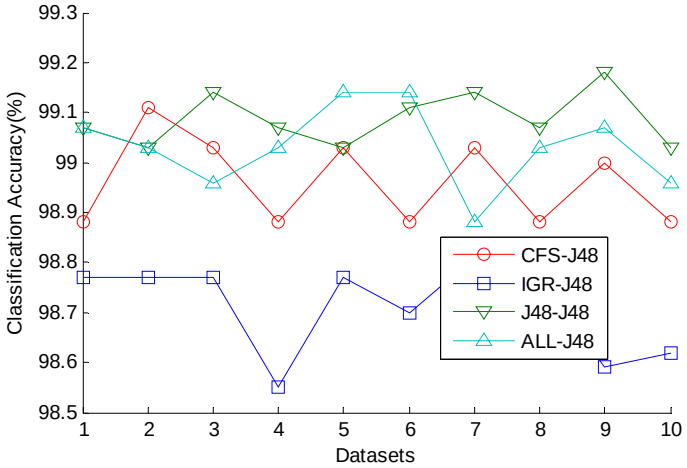

Figure 5.. Classification Accuracies of J48s

likely reason for this is that the eliminated attributes are noise. Moreover, SVM-SVM is based on the wrapper mode. IGR-SVM and CFS-SVM is based on the filter mode. The classification based on wrapper mode is optimized by genetic algorithm. Ii can be drawn from figure 4 that different data has no affect on SVM.

As shown in the figure 5, J48-J48 has the best classification performance. Compared with SVM, J48-J48 is not comprehensively better than ALL-J48 in all datasets. Especially, in the testing dataset2, J48-J48 and ALL-J48 has lower classification accuracy than CFS-J48. But it is no doubt that IGR-J48 has the worst classification performance of them. It can be drawn from figure 5 that different data has a large impact on J48.

It can be concluded from figure 4 and figure 5 that SVM has better stability and better classification performance than J48. IGR, as the front attribute selection, has the worst classification performance.

VI. Conclusion

The intrusion classifier based on multiple attribute selection algorithms has been proposed in this paper. The new system has six combinations with different representative attribute selection algorithms and different classification algorithms. Through comparing with classification performance and real time, advantage or disadvantage of different combinations comes out. It is positive significance for deploying different algorithm combinations based on the concrete context.

In the future, we will try to apply the intrusion classifier into the field of wireless sensor networks. Some core code of intrusion classifier should be simplified. The classifier will be improved to be the next module of the lightweight detection.

Acknowledgment

This work was supported in part by the National High Technology Research and Development Program of China under Grant No. 2011AA010101, the National Natural Science Foundation of China under Grant No. 61103197 and 61073009, the Key Programs for Science and Technology Development of Jilin Province of China under Grant No. 2011ZDGG007, the Youth Foundation of Jilin Province of China under Grant No. 201101035.

References

[1] Sheikhan Mansour, Jadid Zahra, Farrokhi Ali, “Intrusion detection using reduce-size RNN based on feature grouping”, Neural Computing & Applications, vol. 21, no. 6, 2012, pp. 1185-1190.

[2] Sindhu Siva S. Sivatha, Geetha S., Kanan A, “Evolving optimized decision rules for intrusion detection using particle swarm paradigm”, International Journal of System Science, vol.43, no.12, 2012, pp.2334-2350.

[3] Damopoulos Dimitrios, Menesidou Sofia A., Kambourakis Georgios, “Evaluation of anomaly-based IDS for mobile devices using machine learning classifiers”, Security and Communication Networks, vol. 5, no.1, 2012, pp.3-14.

[4] Koc Levent, Mazzuchi Thomas A., Sarkani Shahram, “A network intrusion system based on a Hidden Naïve Basyes multiclass classifier”, Expert System with Applications, vol.39, no.18, 2012, pp.13492-13500.

[5] Chung Yuk Ying, Wahid Noorhaniza, “A hybrid network intrusion detection system using simplified swarm optimization”, Applied Soft Computing, vol.12, no.8, 2012, pp.3014-3022.

[6] Lin Shih-Wei, Ying Kuo-Ching, Chou Yuan, “An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection”, Applied Soft Computing, vol.12, no. 10, 2012, pp.3285-3290.

[7] Badran Khaled, Rockett Peterm, “Multi-class pattern classification using single, multi-dimensional feature-space feature extraction evolved by multi-objective genetic programming and its application to network intrusion detection”, GENETIC PROGRAMMING AND EVOLVABLE MACHINES, vol.13, no.1, 2012, pp. 33-63.

[8] Amiri Fatemeh, Yousefi MohammadMahdi Rezaei, Lucas Caro, “Mutual information-based feature selection for intrusion detection systems”, Computer & Security, vol.30, no.6-7, 2012, pp. 514-524.

[9] Uysal Alper Kursat, Gunal Serkan, “A novel probabilistic feature selection method for text classification”, Knowledge-based System, vol.36, 2012, pp.226-235.

[10] Wang Guangtao, Song Qinbao, Xu Baowen, “Selecting feature subset for high dimensional data via the propositional FOIL rules”, Pattern Recognition, vol.46, no.1, 2013, pp.199-214.

[11] Hongying Zheng, Meiju Hou, Yu Wang, “An Efficient Hybrid Clustering-PSO Algoritthm for Anomaly Intrusion Detection”, Journal of Software, vol. 6, no.12, 2011

[12] Yuesheng Gu, Yongchang Shi, Jianping Wang, “Efficient Intrusion Detection Based on Multiple Neural Network Classifiers with Improved Genetic Algorithm”, Journal of Software, vol.7, no. 7, 2011.

[13] Nabil EL KADHI, Karim HADJAR, Nahla EL ZANT, “A Mobile Agents and Artificial Neural Networks for Intrusion Detection”, Journal of Software, vol. 7, no. 1, 2012.

Wei-wu Ren studies in the College of Computer Science and Technology at Jilin University, and separately gained a bachelor's degree in 2007 and a master's degree in 2010. Now, he is a doctorate candidate in the same university. His research interests are information security, data mining and knowledge representation.

Liang Hu received his Ph.D. degree in Computer Software and Theory from Jilin University in 1999. He is currently a professor in the College of Computer Science and Technology, Jilin University. His research interest covers network security and distributed computing, including related theories, models, and algorithms of PKI/IBE, IDS/IPS, and grid computing.

Kuo Zhao received the B.E degree in Computer Software in 2001 form Jilin University, followed by M.S degree in Computer Architecture in 2004 and Ph.D. in Computer Software and Theory from the same university in 2008. He is currently associate professor in the College of Computer Science and Technology, Jilin University. His research interests are in operating systems, computer networks and information security.

Jianfeng Chu (chujf@jlu.edu.cn) is currently an associate professor in the College of Computer Science and Technology, Jilin University. He received his Ph.D. degree from Jilin University. His research interests include network security, mobile security, cloud security and security of Internet of Things. He is the corresponding author of this paper.

Bing Jia JinLin Province, China. Birthdate: Februrary, 1985. and separately gained a bachelor's degree in 2006 and a master's degree in 2010. Now, Studies for the doctorate the College of Computer Science and Technology at Jilin University. And her research interests are artificial intelligence and active service.