Integrated Intrusion Detection System Using Soft Computing

S Selvakani Kandeeban[1] and Rengan S Rajesh[2]
(Corresponding author: S. Selvakani)
Asst.Prof Department of Computer Applications, Jaya Engineering College[1]
Chennai, Tamilnadu, 602 024, India.
Reader,Dept of CSE, MS University, Tirunelveli, Tamilnadu, 627 009, India[2]
(Email: sselvakani@hotmail.com)
(Received Aug. 1, 2008; revised and accepted Dec. 4, 2008)

Abstract

Intrusion Detection systems are increasingly a key part of system defense. Various approaches to Intrusion Detection are currently being used but they are relatively ineffective. Among the several soft computing paradigms, we investigated genetic algorithms and neural networks to model fast and efficient Intrusion Detection Systems. With the feature selection process proposed it is possible to reduce the number of input features significantly which is very important due to the fact that the Radial Basis Function networks can effectively be prevented from over fitting. The Genetic algorithm employs only the eight most relevant features for each attack category for rule generation. The generated rules signal an attack as well as its category and it is end for training to RBF network. The optimal subset of features combined with the generated rules, can be used to analyze the attacks. Empirical results clearly show that soft computing approach could play a major role for intrusion detection. The model was verified on KDD99 demonstrating higher detection rates than those reported by the state of art while maintaining low false positive rate.

Keywords: Genetic algorithms, information gain, mutual information, radial basis function networks, soft computing

1 Introduction

Along with bringing revolution in communication and information exchange, Internet has also provided greater opportunity for disruption and sabotage of data that was previously considered secure. While we are benefiting from the convenience that the new technology has brought us, computer systems are facing increased number of security threats that originate externally or internally. As malicious intrusions into computer systems have become a growing problem, the need for accurately detecting these intrusions has risen [1, 9, 14, 18]. Despite numerous technological innovations for information assurance, it is still very difficult to protect computer systems. Therefore intrusion detection is becoming an increasingly important technology that monitors network traffic and identifies, preferably in real time, unauthorized use, misuse and abuse of computer systems. A number of approaches based on computing have been proposed for detecting network intrusions. The guiding principle of soft computing is exploiting the tolerance of imprecision, uncertainty, partial robustness and low solution cost. Soft computing includes many theories such as fuzzy logic, neural networks, artificial intelligence, information and probabilistic reasoning and genetic algorithms.

When used for intrusion detection, soft computing techniques are often used in conjunction with rule-based expert systems where the knowledge is usually in the form of if-then rules. Despite different soft computing based approaches having been proposed in recent years [16], the possibilities of using the techniques for intrusion detection are still under utilized. In our approach we consider Intrusion detection as a data analysis process. Network behaviors can be categorized into normal and abnormal. Due to the sheer volume of real network traffic, both in the amount of records and in the number of features, it is extremely difficult to process all the traffic information before making decisions. We need to extract the most important data that can be used to efficiently detect network attacks. We use information theory to identify the most relevant features to be used [15]. In this work the initial point is the extraction of the most important piece of information that can be deployed for efficient detection of attacks in order to cope with this problem. Our idea is to achieve high detection rate by introducing high level of generality when deploying the subset of the most important features of the dataset. As this also results in high false positive rate, we deploy additional set of rules in order to recheck the decision of the rule set for detecting attacks. We deploy genetic algorithm (GA) approach for

offline training of the rules for classifying different types of intrusions. Genetic Algorithm field is one of the upcoming fields in computer security and has only recently been recognized as having potential in the Intrusion Detection field. Neural Networks have been actively applied to IDS. To apply neural networks to real world problem successfully, it is very important to determine the number of hidden nodes in the given problem, because performance hinges upon the structure of the neural networks. Hence we use RBF for learning the rules. We examine the proposed method through experiments with real data and compare the results with those of other methods. The aim here is to develop an Intrusion Detection System which adapts to the environment. Evolution and learning are the two most fundamental process of adaptation. Since learning through neural network is a complex, time consuming process, the connection between learning and evolution can be used to decrease the complexity of the problem. The rest of the paper is organized as follows: Section 2 presents an overview of related works. Section 3 gives overview on attacks within the KDD data set and proposes the deficiencies of the data set. Section 4 gives overview on Genetic Algorithms and the benefits of its using in intrusion detection field. Section 5 discusses the detection rate of our algorithm when applied to the KDD 99 data.

2 Related Works

A. Chittur extended their idea by using GA for anomaly detection [3]. Random numbers were generated using GA. Random numbers were generated using GA. A threshold value was established and any certainty value exceeding this threshold value was classified as a malicious attack. The experimental result showed that GA successfully generated an empirical behavior model from training data. The biggest limitation of this model was the difficulty of establishing the threshold value which might lead to detect novel or unknown attacks. Gomez et al. [4] proposed a linear representation scheme for evolving fuzzy rules using the concept of complete binary tree structure. GA is used to generate genetic operators for producing useful and minimal structure modification to the fuzzy expression tree represented by chromosomes. The biggest drawback of the proposed approach was that the training was time consuming. Liao and Vemuri used the K-nearest Vector Machine [12] for profiling computer programs. The KNN classifier was employed with an interesting analogy between classifying text documents and detecting intrusion using the sequences of system calls.

Wang et al. used the evolutionary algorithm [17] for discovering neural networks for intrusion detection. The connections of the network and its weights were encoded with binary bits and evolved simultaneously. Their detection system was evaluated with www log data and showed an accuracy rate of 95%. However, in their experiment, they used their own data set rather than a public bench mark dataset. Hofmann et al. proposed [8] the evolutionary learning of radial basis function networks for intrusion detection. They targeted a network based IDS. Their evolutionary algorithm performed two tasks simultaneously selecting the optimal feature set and learning the RBFN. The binary bits system was used to encode the 137 possible features of the network packet headers and three components of the RBFN, including the type of basis function, the number of hidden neurons, and the number of training epochs. In the experiments with the network audit data set, the RBFN optimized with the evolutionary algorithm outperformed the normal MLP and the normal RBFN. Gonzalez et al. proposed an intrusion detection technique based on evolutionary generated fuzzy rules [6]. The conditional part of the fuzzy detection rules was encoded with binary bits and fitness was evaluated using two factors: the accuracy and the coverage of the rule. The performance was compared to the methods of different genetic algorithms and without the fuzziness of rules using two network audit datasets their own wireless dataset and the knowledge discovery and data mining cup 99 data set.

3 KDD Data Set - Issues and Solutions

Learning algorithms have a training phase where they mathematically learn the patterns in the input data set. The input data set is also called the training set which should contain sufficient and representative instances of the patterns being discovered. A data set instance is composed of features.

In order to promote the comparison of advanced research in the area of intrusion detection, the Lincoln Laboratory at MIT, under DARPA sponsorship, conducted the 1998 and 1999 evaluation of intrusion detection [13]. Based on binary TCP dump data provided by DARPA evaluation, millions of connection statistics are collected and generated to form the training and test data in the classifier Learning contest organized in [13].

The data set contains 5,000,000 network connection records. A connection is a sequence of TCP packets starting and ending at some well defined moments of time, between which data flows from source IP to a target IP. The training portion of the dataset "kdd-10-percent" contains 494,021 connections of which 20% are normal. The testing data set "corrected" provides a data set with a significantly different statistical distribution that the training data set and contains an additional 14 attacks. It contains 311,029 connections of which 60,593 are normal.

KDD data set is comprised of records. Each record in the data set consists of 41 features [10] where 38 are numeric and 3 are symbolic defined to characterize individual TCP sessions. Each record is also labeled i.e. the information whether it represents an attack or a normal connection is also provided.

As the first step in our work, in order to cope with the

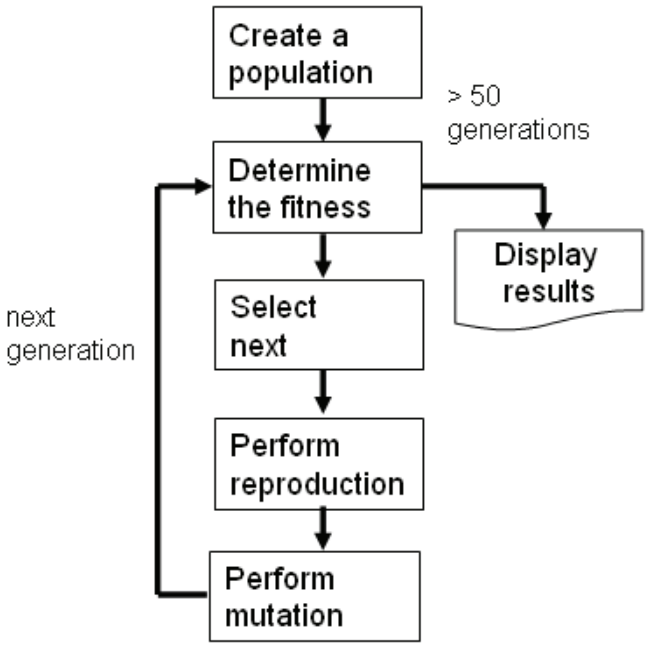

Figure 1: Genetic algorithm flow

speed problem mentioned above, we have used the results obtained in our previous work [15] where we deployed Information Gain based Mutual Information, in order to extract the most relevant features of the data. In this way the total amount of data to be processed is highly reduced. As an important benefit of this arises the high speed of training the system thus providing high refreshing rate of the rule set.

Subsequently, these features are used to form rules for detecting various types of intrusions. This permits the introduction of higher level of generality and thus higher detection rates. The problem that arises with discarding features is a certain increase of false alarm rate.

4 GA Approach

Genetic Algorithms are search algorithms based on the principles of natural selection and genetics. GA evolves a population of initial individuals to a population of high quality individuals, where each individual represents a solution to the problem to be solved. Each individual is called chromosome and is composed of predetermined number of genes.

The quality of each rule is measured by a fitness function as the quantitative representation of each rule's adaptation. The flow is presented in Figure 1.

The procedure starts from an initial population of randomly generated individuals. Then the population is evolved for a number of generations while gradually improving the qualities of the individuals in the sense of increasing the fitness value as the measure of quality.

During each generation, three basic genetic operators are sequentially applied to each individual i.e. Selection, Cross over and Mutation.

Those chromosomes with a higher fitness value are more likely to reproduce offspring. If the new generation contains a solution that produces an output that is close enough or equal to the desired answer then the problem has been solved. If this is not the case, the new generation will go through the same process. This will continue until a solution is reached.

5 Evaluation and Results

The implemented system for intrusion detection consists of a combination of three different parts. The first block represents a feature extraction by using Information gain [15]. In the continuation the next block represents a rule based system using Genetic Algorithm for known attacks. The last block is the detection of unknown attacks using RBF networks.

The raw data from the KDD 99 is first partitioned into four groups (input data set), DoS attack set, Probe attack set, U2R attack set and R2L attack set. For each attack set different connection record feature set are selected as attributes. The goal of this experiment is to generate a rule which can detect whether a connection is an attack or normal. Each of them forms as "if (condition) then attack" where condition is made up of attributes which belong to feature subset and their values. The proposed method is implemented using the JAVA Language.

A fundamental step in the design of a GA is the compatible representation of the problem. For the rule generation problem, two representations have been used, each individual in the population (a chromosome) encodes a single rule or each chromosome encodes a set of rules. Both approaches have their pros and cons. We adopt each individual in the population encodes a single rule in this work because of its efficiency resulting from the smaller individuals generated by the approach.

A rule consists of two parts: the antecedent and the consequent. In general, the antecedent part is a conjunction of a number of attribute values. Each categorical attribute in the conjunction is represented by k bits, where k is the number of possible values for the attribute. If a particular value is present then the corresponding bit is set to 1; otherwise, it is zero. Multiple bits may be set to 1 representing a logical OR among the corresponding attribute values. Multiple attributes in the antecedent are chained by the logical AND operator. Numeric attributes are represented by k bits, where k is the number of bins into which the attribute is categorized using equi-width binning.

If the value of the attribute lies within a bin, then the corresponding bit is set to 1; otherwise, it is zero. The consequent part consists of the attack category attribute and its value only. It is represented by m bits, where m is the number of attack categories. If an attack cate-

gory is present then the corresponding bit is set to 1 and other bits are zero. The consequent part is encoded in the genome of the individual and participates in the evolution process. Our algorithm is run m times, once for each possible value (attack category), generating the best rules for each attack category.

The initial selection for the rules (or individuals in the population) can be done randomly or by using the training dataset to select rules that cover many records. In subsequent iterations of the GA, selection is done based on the fitness value of the rules.

The crossover operation is an important step in a GA. One-point and two-point crossover operations are the most commonly used types. In the one-point crossover operation, the two rules taking part in the operation are cut at a randomly selected point and their left (or right) parts are interchanged. This crossover operation is sensitive to the position of the attributes within the chromosome. For example, the left most and right most attributes are highly unlikely to go in the same partition. We use the two-point crossover operation, in which two cut points are randomly selected in the two rules and the part within the cut point is interchanged. Two-point crossover is independent of the position of the attributes in the rule. The selection of chromosomes that take part in the crossover operation is done by the roulette wheel strategy. The population size is kept constant from one iteration to the next.

Mutation is random flipping of the bits in the rule. It helps avoid getting stuck in local maxima in the search. We use single bit mutation with a low flipping probability.

Each chromosome, representing one learning rule, was evaluated. To evaluate a chromosome, an appropriately sized network was configured for each of the 20 tasks. The following procedure was conducted for each task. For each epoch, the network was shown all the training patterns, and the weights were updated according to the encoded learning rule. The absolute values of connection strengths were capped at 20 to prevent runaway learning rules. The network was presented with each pattern once more, and its outputs were recorded. If the desired and actual outputs were on opposite sides of 0.5, the response was counted as an error.

An individual (a chromosome) of each population consisted of genes, where each gene represented a certain feature and its values represented the value of the feature. Each GA is trained during 300 generations where in each generation 100 worst performed individuals are replaced with the newly generated ones.

The performance is measured by the fitness function as

$$\text{Fitness} = 3 * (\alpha/A) - 2 * (\beta/B).$$

Where

$\alpha \rightarrow$ number of correctly classified connections;

$\beta \rightarrow$ number of incorrectly classified connections;

$A \rightarrow$ total number of attacks;

$B \rightarrow$ total number of normal connections.

High detection rate and low false positive rate result in high fitness value. On the other side low detection rate and high false positive rate results in low fitness value. Greater coefficients at the side of the detection rate give a certain advantage of the detection over false positive rate. The coefficients were chosen after having performed a number of experiments where they assumed different values. This function was used for its simplicity and ease of interpretation.

- The fitness of a chromosome was taken as its average fitness over all twenty tasks, and chromosomes were probabilistically selected for inclusion in the next generation based on their cumulative fitness over generations. The selection mechanism was roulette selection with elitism (that is, the most-fit chromosome was always included in the next generation).

- After the 500th generation, the best chromosome was selected, and the learning rule was encoded using it. The fitness of the learning rule was tested on the 10 datasets that were held back for the testing task. The same process was repeated for ten epochs and the results were analyzed. The above process was repeated with different parameters of genetic algorithm. A two point cross-over and elitist selection was used. The cross over rate was varied from 50% to 80% with an increment of 5% in every step. The mutation rate was varied from 1% to 5% with an increment of 0.5%. The sample rules are the following ones:

  If (duration $\leq$2 & num_failed_login $>$ 5)
  Then guess_passwd.
  If(Protocol_type = icmp & src_bytes $>$ 300)
  Then Smurf.
  If(Service = ftp_data & flag = REJ)
  Then ipsweep.
  If(protocol_type = udp & src_bytes $\leq$6)
  Then Satan
  $\vdots$
  Default: normal.

  Let us now analyze the results obtained from the experiments conducted as explained above. As discussed, we tried the proposed algorithm with different sets of parameter values. The best fitness was 92.4% which was achieved with 55% cross-over rate and 1% mutation rate.

The result of the training process is certain number of best performed rules. We have performed experiments using 50, 75 and 100 best performed rules for detecting attacks.

Table 1: Detection rate

Number of rules	Detection Rate			
	Dos	U2R	R2L	Probe
50	86.7%	79.2%	81.2%	86.1%
75	81.4%	71.3%	75.4%	83.4%
100	78.3%	67%	71%	82.5%

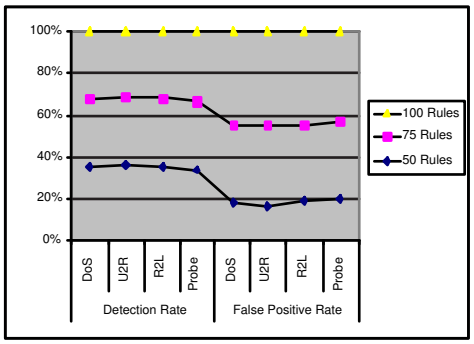

Figure 2: ROC curve for ID performance

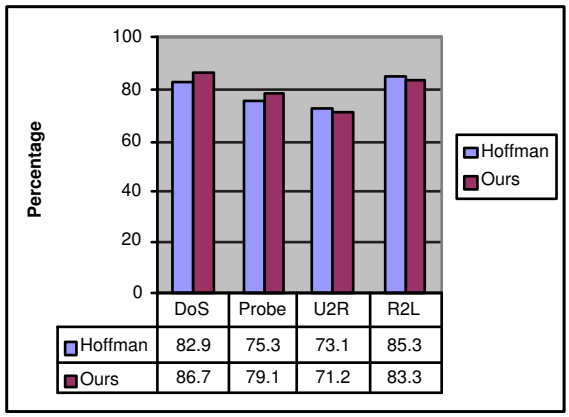

Figure 3: Performance comparison

From Table 1 believing that our system outperforms the best performed model reported in literature. Moreover, our previous statement of reducing false positive rate when deploying additional rule systems for detecting Dos attacks and normal connections is conformed, as the false positive rate has decreased in each of the cases.

Ten kinds of network attacks are included in the training set, namely back, land, Neptune, Pod, smurf, teardrop, ipsweep, portsweep, buffer_overflow and guess_passwd. Fifteen kinds of network attacks are included in the testing data set namely perl, xlock, mailbomb,UDPstrom, saint, xlock, back, land, Neptune, pod, smurf, teardrop, ipsweep, portsweep, bufferoverflow and guess_passwd. These fifteen attacks are also extracted from the KDD Testing data set. The test data set is similar with the training data set. The only differences are that the test data set includes some unknown attacks not accruing in the training data set. Detection of attacks involved in the test data set and not occurring in the training data set assesses the potential ability to detect novel attacks.

To show the performance visually, the receiver operating characteristic (ROC) curve, which is a traditional way to represent the performance of the classifier, is used. Figure 2 depicts the ROC curve which illustrates the intrusion detection performance of the proposed method. The DR increases as the false alarm rate does the same. The DR is close to 95 when the false alarm rate is 1.9% to 2%. However the false alarm rate is close to 1%, the DR is only about 70%. The reason is because the number of rules in the rule base is different for each run.

The Performance of ours is compared to that of Hoffman GA Rules for Intrusion Detection in Figure 3. Ours outperforms the winning result for DoS and Probe and is not significantly worse for R2L and U2R attack categories. These results highlight the advantage of using reduced and relevant feature set. The performance is comparable and at the same time bring more efficient.

6 Conclusion

In this work a combination of GA based ID for detecting different types of attacks was introduced. As we use only nine features to describe the data, its time of training is considerably reduced, thus providing high refreshing rate of the rule set. However the detection rate is not good for some runs because of the selection of cross over and mutation points in corresponding operations is random. The linear structure of the rule makes the detection process efficient in real time processing of the traffic data. The evaluation of our approach showed that the hybrid method of using discrete and continuous features can achieve a better detection rate of network attacks. In order to increase the detection rate, we select the features that are appropriate for each type of network attacks. That is also an added advantage.

Further enhancements should be made by the rule learning technique using Radial Basis Network for detecting any unknown attacks.

References

[1] T. Bhaskar, N. Kamath B, and S. D. Moitra, "A hybrid model for network security systems: Integrating intrusion detection system with survivability," *International Journal of Network Security*, vol. 7, no. 2, pp. 249-260, 2008.

[2] Y. Bouzida, and F. Cuppens, "Detecting known an novel network intrusion," *IFIP/SEC 2006, International Information Security Conference*, pp. 123-129, Karlstad University, Sweeden, May 2006.

[3] A. Chittur, *Model Generation for an Intrusion Detection System Using Genetic Algorithms*, High school Honors Thesis, accessed in 2006. (http://www1.cs.columbia.edu/ids/publications/gaids-thesis01.pdf)

[4] J. Gomez, D. Dasgupta, D. Nasaroui, and F. Gonzalez, "Complete expression trees for evolving fuzzy classifiers system with genetic algorithms and applications to network intrusion detection," *Proceedings of NAFIPS-FLINT Joint Conference*, pp. 469-474, New Orleans, LA, June 2002.

[5] R. H. Gong, M. Zulkernine, and P. Anolmaesumi, "A software implementation of a genetic algorithm based approach to network intrusion detection," *Proceedings of the SNPD/SAWN' 05*, pp.19-27, Aug. 2005.

[6] F. Gonzalez, J. Gomez, M. Kaniganti, and D. Dasgupta, "An evolutionary approach to generate fuzzy anomaly signatures," *Proceedings 4th Annual IEEE Information Assurance workshop*, pp. 251-259, West point, NY, June 2003.

[7] C. Grosan, A. Abraham, and M. Chis, "Computational intelligence for light weight Intrusion Detection systems," *International Conference on Applied Computing, IADIS' 06*, pp. 538-542, San Sebastian, Spain, May 2006.

[8] A. Hofmann, and B. Sick, "Evolutionary optimization of radial basis function networks for intrusion detection," *Proceedings of nternational Joint Conference Neural Networks*, vol 1, pp. 415-420, Portland, OR, July 2003.

[9] P. Kabiri and A. Ghorbani, "Research on intrusion detection and response: A survey," *International Journal of Network Security*, vol. 1, no. 2, pp. 84-102, 2005.

[10] KDD cup 1999 data. (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html)

[11] P. Lascov, P. Dussel, C. Schafer, and K. Riek, "Learning intrusion detection: Supervised or unsupervised?," *CIAP: International Conference on Image Analysis and Processing*, vol. 3617, pp. 50-57, cagliari, Italy, Sep. 2005.

[12] Y. Liao, and V. R. Vemuri, "Use of k-nearest neighbour classifier for intrusion detection," *Computer Security*, vol. 21, no. 5, pp. 439-448, Oct. 2002.

[13] MIT Lincoln Laboratory DARPA Intrusion Detection Evaluation. (http://www.ll.mit.edu/IST/ideval/index.html)

[14] I. V. Onut and A. A. Ghorbani, "A feature classification scheme for network intrusion detection," *International Journal of Network Security*, vol. 5, no. 1, pp. 1-15, 2007.

[15] S. Selvakani, and R. S. Rajesh, "Improving ID performance using GA and NN," *International Journal of Computer Aided Engineering and Technology*, vol. 13, no. 1, pp. 81-93, 2008.

[16] N. Srinivasan and V. Vaidehi, "Performance analysis of soft computing based anomaly detectors," *International Journal of Network Security*, vol. 7, no. 3, pp. 436-447, 2008.

[17] L. Wang, G. Yu, G. Wang, and D. Wang, "Method of evolutionary neural network based intrusion detection," *Proceeding of International Conference Infotech and Info-net*, vol. 5, pp. 13-18, Beijing, China, Oct. 2001.

[18] Z. Zhang, H. Shen, Y. Sang, "An observation-centric analysis on the modeling of anomaly-based intrusion detection," *International Journal of Network Security*, vol. 4, no. 3, pp. 292-305, 2007.

S. Selvakani is an Assistant Professor of MCA Department at Jaya Engineering College, Chennai. She received her MCA degree from Manonmanium Sundaranar University and M.Phil degree from Madurai Kamaraj University. She has presented 4 papers in National Conference and 1 paper in international conference. She has published 1 paper in National journal and 5 papers in International Journal. She is currently pursuing her Ph.D degree in Network Security.

R. S Rajesh received his B.E and M.E degrees in Electronics and Communication Engineering from Madurai Kamaraj University, Madurai, India in the year 1988 and 1989 respectively, and completed his Ph.D in Computer Science and Engineering from Manonmaniam Sundaranar University in the year 2004.