





Empirical Analysis of Data Mining Techniques in Network Intrusion Detection Systems

Reza Soufizadeh¹ (✉)  and Jamshid Bagherzadeh² 

¹ Department of Computer Engineering, Science and Research Branch,
Islamic Azad University, West Azerbaijan, Iran

R.Soufizadeh@iaurmia.ac.ir

² Department of Computer Engineering, Urmia University, Urmia, Iran
J.Bagherzadeh@urmia.ac.ir

Abstract. Computer networks have an essential role in modern societies which are developing extensively. Considering that the main goal of the attackers is the ability to access a huge amount of information, Intrusion detection techniques have been attracted attention to the researchers and they believe that to proffer an approach that has an optimization rate both in timing and performance recognition. Moreover, it can be also implemented in commercial devices. A complete analysis of the latest researches in anomaly detection with a high recognition rate of 98% and 2% false positive can be reported. Despite the high rate of attack detection in academic researches, looking at industry solutions that are commercially produced, fewer products can be found that implement smart methods on devices. However, cybersecurity engineers still do not believe in the performance of these new technologies. In order to find out the reason for this contradiction, NSL-KDD and KDDCUP99 Data sets with some machine learning approaches will be evaluated and the results will be compared with previous related works in this paper.

Keywords: Network Intrusion Detection System · Classification methods · Data mining and Machine learning · Feature selection

1 Introduction

Computer networks have a major role in today's modern world and they are developing and becoming inclusive rapidly. At the same time, ensuring their security, maintenance and stability require a high cost. Since the main purpose of attacks is to reach the high amount of information, intrusion detection techniques, have attracted researchers attention. They attempt to find a way that is efficient from both aspects of time and detection ability, and at the same time the technique should be capable of being implemented in network security devices. Network attacks as a group of destructive activities are known for fragmentation, denial and destruction of the information and services in computer networks. For example, network attacks are viruses attached to e-mails, system's probe for collecting information, internet worms, unauthorized use of a system and denial of services with abuse of system's attributes or exploiting a bug in software in order to change the system's information.

An intrusion detection system (IDS) can be either a device or a software application by which a network or a system is monitored for malicious activity or policy violations. Any malicious activity or violation is typically reported either to an administrator or collected centrally using a security information and event management (SIEM) system. A SIEM system combines outputs from multiple sources, and then uses alarm filtering. Intrusion detection typically refers to tools for detecting efforts which want to unauthorized access to a system or to decline its efficiency. In other words, these systems with checking the saved information of user’s loggings do not permit to any unauthorized login to the system and meanwhile they detect the users’ activities while they are doing something on a system in order to inform the system’s manager if there is an unauthorized activity by a user. A simple model for network intrusion detection system has shown in Fig. 1:

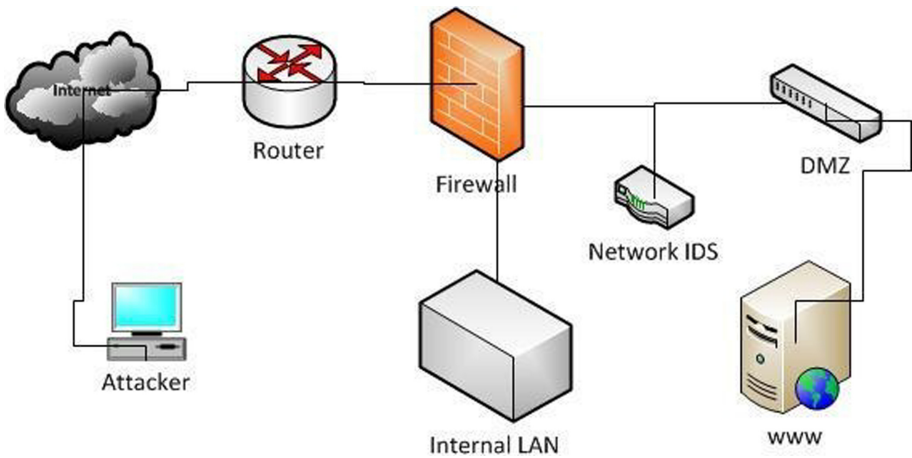


Fig. 1. A simple model of exposure IDS in computer networks

2 Network Intrusion Detection Systems

Network intrusion detection systems (NIDS) like other network equipment are developing in attacks’ detection aspect. For a long time, intrusion detection systems have been focusing on anomaly detection and misuse detection. Meanwhile, commercial manufacturers concentrate highly on misuse detection for high level of detection ability and high amount of precision. Anomaly detection is being developed in academic researches for the existence of high level of theoretical background. This method as a general analysis features like; CPU consumption, input and output, traffic network card, number of file access, user’s identity, machines that a user want to access, all of the opened files, read pages and page fault. Then with being far from the threshold, by using statistical or intelligent techniques, it will be detected as anomaly [1]. In misuse detection methods, patterns that are clear in data course are first encoded then corresponded with intrusive procedures like special signatures [2]. At the same time, anomaly detection, a model of data flow, is being monitored by statistical analysis to detect whether in normal situations, intrusive procedures, abnormal traffic, and an unusual activity happened as intrusion or

not [1]. In addition, it is difficult to recognize signatures that include different types of possible attacks. All the mistakes in detecting these signatures cause the increase of false alarm rate and decrease of detection technique's efficiency. Therefore, techniques which are based on rules can be used. Thus, the security expert can form the policies as rules then it is corresponded with data flow model. It is imperative that the methods based on rules in corresponding patterns be updated by security experts [2].

3 KDDCUP99 and NSL-KDD Datasets

Different data sets with various classifications have been presented for attacks up to now, but the [3] classification seems to be more complicated and more complete than the others and at the same time includes the whole qualities and capabilities of other classifiers. If there is a better description of attacks, the detection of them can be easily done by machine learning techniques. Since 1999, KDDCUP99 data set has been used in order to evaluate the anomaly detection method widely. This data set was prepared at Lincoln laboratory of MIT University by Stolfo et al. during 7 weeks with approximately 5 million records of data and the capacity of 4 GB in which each record had 100 bites capacity; this data set also constituted 41 features [4].

As regards with having a comprehensive analysis of the recent process in anomaly detection and according to previously reported researches which has been mentioned above, the highest detection rate of 98% and false detection rate of 2% can be obtained [5]. Despite highest rate of attack detection in academic researches, you can't see any machine learning methods in produced commercial devices. That's the reason, cyber security equipment manufacturers do not believe to efficiency of recently introduced technologies. In order to find out the reason of this contradiction, A.A. Ghorbani et al. [6] investigated the details of accomplished studies in anomaly detection domain and its different aspects, including: training, learning, testing and evaluation of data sets with variety methods. Their studies reveal that there are intrinsic problems in KDDCUP99 data set. Nevertheless, most of the researchers use this data set which is one of the prevalent data sets for anomaly detection and obtain unreliable results for ages. The first shortcoming of KDDCUP99 data set is the large amounts of data redundancy.

As regards with analyzing, training and testing data sets, it can be realized that nearly 78% and 75% of records of these sets are duplicated [6]. This large amount of data redundancy in the training set causes the machine learning algorithms don't have a good performance. As a result, having duplicated records in both testing and training sets has been reported a high percentage of detection by previous researchers in this area. While studying different machine learning algorithms and randomly selected instances from data sets as mentioned before, a high detection rate of 98% can be obtained. This amount is declined to approximately 86% in the worst conditions. A.A. Ghorbani et al. in [6] their research, by presenting KDDCUP99 problems acknowledged that the evaluated results in this area are unreliable. On the other hand, the existence of redundant, duplicated and repeated records in both testing and training tables is harmful and in reported papers the detection rates of these attacks are lower than other ones. Nevertheless, there is only a few numbers of such attacks in both tables and they do not follow a normal distribution. Thus, as the first step the redundancy of the training and testing data set records are eliminated and then the train records are eliminated which are repeated in the test table.

A new data set is presented as NSL-KDD in [8]. Although this new data set does not have the above mentioned problems, it still has the problems asserted by McHugh [7].

4 Related Works

Nowadays with the extensive development of computer networks and the rapid increase of special applications running in these networks, the importance of the security of these networks is being concerned. During the last decade, misuse and anomaly detections have been more concerned. The researchers about overcoming the flaws of misuse detection in novel attacks, and KDDCUP99 data set is highly being used for evaluation systems. For a long time, researches on intrusion detection range had been concentrated on anomaly and misuse detections. Since misuse detection is concentrated by commercial manufacturers for high level of detection ability and high amount of precision, anomaly detection is developing for the existence of high level of theoretical background in academic researches.

4.1 Naïve Bayes Method in Anomaly Detection

Conditional probability $P(H|E)$ is used to compute the probability of H given E. H can be sampled as a column feature vector and can be considered as $X = x_1, x_2, \dots$. We calculate: $P(X|class = Normal) \cdot P(Normal)$ and $P(X|class = Attack) \cdot P(Attack)$, each part that becomes maximum, indicates that input data is Normal or Attack respectively. Adebayo et al. [9] has eliminated these features with using fuzzy methods but he has not given a clear explanation of how he did it: 0, 1, 8, 14, 15, 16, 17, 18, 19, 20, 21, 36 features from their test and carried out their evaluations based on only 22 features and used Naïve Bayes method with 5924 training data and 12130 test data, and finally the results were the same as those obtained from the whole features equal to 96.67%. Ben Amor et al. [10] for *DoS*, *U2R*, *R2L* and *Probe* attacks as well as for the normality of input packets using Naïve Bayes method obtained the accuracy of 96.38%, 11.84%, 7.11%, 78.18% and 96.64% respectively. At the same time the precision of 98.48% and 89.75% was reported for normal and abnormal detections respectively.

4.2 Decision Trees Method in Anomaly Detection

In artificial intelligence, trees are used for different concepts such as: sentences structures, equations, game modes, and so on. Decision trees learning is a way for approximation of the objective functions of discrete values. This method, which is resistant to noise of data, is able to learn the disjunction predicate conjunction. Pachghare et al. [11] detected the level of packet's normality about 99% without any preprocessing only by using decision trees and 1000 instances. In [13], with using "Feature Selection" technique and "InfoGain" method, the accuracy rate of 95% was obtained.

4.3 Support Vector Machine Method in Anomaly Detection

The main idea of the support- vector machines, [12, 13] is to increase the samples size as they can be separated. Hence, despite the fact that there is a common process in order to reduce the dimensions in the support vector machines, in reality the dimensions increase. The aim is to find a very dimensions, it may seem excessive as a volume). Teng et al.

[15] using the fuzzy and SVM methods and also dividing test dataset and train dataset to three groups performed their tests based on TCP, UDP, ICMP protocols and at the end they obtained 82.5% accuracy rate for a Single SVM and 91.2% of accuracy for a Multi SVM. In [15] article, Rung-Ching et al. obtained 89.13% of accuracy using SVM and Rough Set methods.

4.4 Artificial Neural Networks Method in Anomaly Detection

Multilayer perceptron (MLP) [12] is one of the most common algorithms being used in neural networks classification. Researchers use multilayer perceptron for detection of the attacks in KDDCUP99 data set [16]. Their structure consists of Feed-Forward, three-Layer neural networks: an input layer, a hidden layer and an output layer. Unipolar sigmoid transfer functions for each neuron in both hidden and output layers are used with deviation value of 1. The applied detection algorithm is a random descending gradient with the mean square error function. As a whole, there are 41 neurons in the input layer (pattern with 41 input features) and 5 neurons (one for each group) in the output layer. The reported results show that 88.7% of attacks are *probe*, 97.2% are *DoS*, 13.2% are *U2R* and 5.6% of attacks are *R2L* [16]. In [17], Abdulkader et al. using neural networks for some special *DoS* attacks with 24 neurons and a hidden layer, obtained 91,42% detection rate with 8,57% false detection rate. Their test revealed that even if they increased the number of neurons, the above ratios would not change. While Mukhopadhyay et al. used the back propagation neural network [18] with all 41 features; they used corrected data set as learned and test. As a result, from 311030 records of this data set, they used 217720 records for train and 46655 records as test and finally they reported 95.6% detection rate with 4.4% false detection rate.

5 Evaluation Made by Intelligence Algorithms on KDDCUP99 and NSL-KDD Datasets

As already mentioned, different tables have been extracted from KDDCUP99. Generally, in the published papers, random samples are used from `kddcupdata10percent` table, for training and testing, which finally give unreliable results. In this research, first of all the tables are selected using KDDCUP99 data set for evaluation and then they are compared with similar related works. In the next step, evaluations are done based on NSL-KDD Data Set as follows and finally the results are compared.

5.1 Preprocessing and Analysis of Various Methods on KDDCUP99 Data Set

First of all, from KDDCUP99 data set 10% of the corrected table is selected randomly as testing data with 17 novel attacks, and 10% of `kddcup.data_10_percent` table as training data. Analyzing the information in the tables with SQL Servers facilities (see Table 1), it can be clearly seen that `num_outbound_cmds` feature has the value of zero in all rows. Therefore, this feature is not used in our computations using machine learning techniques and the following results can be obtained:

Table 1. Random sample selection from KDDCUP99

Instances to Test	Instances to Train	R2L	U2R	DoS	Probe	Normal
31103	49402	988	3	23627	809	5676

We evaluated various methods on KDDCUP99 and compared them with [13–16] which are shown in Table 2 and Table 3.

Table 2. Comparison of accuracy various methods on KDDCUP99

Method	Accuracy	Attack	Normal
Ref [13] Decision tree feature selection	95.02%	–	–
Ref [16] Hybrid methods with 41 feature	95.06%	–	–
Decision Tree with 40 feature	96.32%	95.7%	99%
Naïve Bayes with 40 feature	96.42%	96%	97%
Neural Networks with 40 feature	96.56%	–	–
Ref [14] SVM	91.2%	–	–
Ref [15] SVM Feature Selection	95.65%	–	–
Single SVM with 40 feature	95.71%	–	–

Table 3. Analysis of various methods on KDDCUP99

Category of attacks	Ref [10] Naïve Bayes with 41 features	Naïve Bayes 40 features	Decision tree 40 features	Ref [16] Hybrid methods 41 features	Neural Networks 40 Features	Multi SVM 40 features	Naïve Bayes + DT 41 features
DoS	96.38%	99.4%	99.5%	97.2%	97.2%	99.81%	96.25%
U2R	11.84%	75.9%	66.7%	13.2%	0	0	23.68%
R2L	7.11%	0,09%	0,06%	5.6%	0	0	0,014%
Probe	78.18%	93.7%	99%	88.7%	95.67%	94.8%	60.37%
Normal	96.64%	95.8%	99.4%	–	98.90%	99.43%	94.12%

5.2 Preprocessing and Analysis of Various Methods on NSL-KDD Data Set

According to invalid results mentioned before, in order to obtain reliable and acceptable results, NSL-KDD data set will be used in this research. Generally, for obtaining high percentages in researches by using this data set, only the training table is used and unreliable results are obtained. For this reason in this research, from NSL-KDD data set

50% of records are extracted from two NSL-Train and NSL-Test tables randomly with an appropriate distribution of *Protocol*, *Service* and *Flag* features, by using a simple SQL command, then we will compare the results of learning machines with related works. When different researches are reviewed, it can be realized that the only valid and reliable research that corroborates our method of study is the research of A.A. Ghorbani et al. in [6]. According to the analysis on the tables in SQLServer, it is revealed that the *num_outbound_cmds* feature, in both tables has the value of zero for all rows. The nature of this field is used in ftp and has nothing to do with IDS. Accordingly, this feature is not used in our computations using machine learning methods. The results are shown by Table 4 and Table 5:

Table 4. Analysis of various methods on NSL-KDD

Category of attacks	Naïve Bayes 40 features	Decision tree 40 features	Neural networks 40 features	Multi SVM 40 features	Naïve Bayes + DT 41 features
DoS	70%	80%	72%	70%	70%
U2R	14%	0,08%	0.01%	0	0.045%
R2L	17.5%	16.2%	0.01%	0.09%	0.065%
Probe	86.5%	66.8%	49.7%	50%	65.2%
Normal	91.8%	98.5%			
Attack	71.9%	75.8%			

Table 5. Comparison with 40 features and Ref [6]

Methods	Ref [6] All features	With 40 features
Naïve Bayes	76.56%	78.24%
Decision Tree	81.05%	83.78%
Multi-Layer	77.41%	78.4%
Perceptron		
SVM	69.52%	70.8%

It can be concluded from Table 2, Table 3 and Table 4, Table 5 that:

- 1- The Naïve Bayes classification method for the detection of *U2R* and *R2L* *R2L* and *Probe* attacks is better than other approaches.
- 2- The Decision Trees classification method for the detection of *DoS* and *Probe* attacks is better than other approaches.
- 3- The Neural Networks classification method for the detection of *DoS* attacks is better than other approaches.
- 4- The Support Vector Machine classification method for the detection of *Normal* packets is better than other approaches.
- 5- The accuracy of Neural Networks for indicating of Normal/Attack is better than other approaches.

6 Feature Selection

Some studies on KDDCUP99 NSL-KDD data sets' showed researchers among feature selection techniques, select features that are important in the computation of accuracy and false positive and false negative detection. Moreover, they select the features most relevant to each other. Indeed, unnecessary features that decrease accuracy are ignored. These techniques increase the performance and reduce the time compared to normal situation (without selecting feature). InfoGain method is used for selection of features. Using this method has some problems in some attacks which will be discussed later. In this research, InfoGain method is used for selection of the most relevant features and then based on Naïve Bayes.

6.1 InfoGain

Suppose S is the set of labels with the corresponding labels and there are m classes and the sample s_i content from class I and s the number of samples in the train set. The expected information needed to classify a given set is calculated according to the following formula [19]:

$$I(s_1, s_2, \dots, s_n) = - \sum_{i=1}^m \frac{s_i}{s} \log_2 \frac{s_i}{s} \quad (1)$$

The property F with values of $\{f_1, f_2, \dots, f_v\}$ can be added to the training set inside v with subsets $\{S_1, S_2, \dots, S_v\}$ so that S_j is a subset which has the value f_j for the attribute F . Furthermore, S_j is include S_{ij} samples of class i . The entropy of the attribute F is obtained by the following formula:

$$E(F) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{s} * I(S_{1j}, \dots, S_{mj}) \quad (2)$$

Therefore:

$$\text{InfoGain}(F) = I(s_1, s_2, \dots, s_n) - E(F) \quad (3)$$

In this case, If we accomplish InfoGain algorithm on NSL-KDD data set we obtain this features: *Duration, service, src-bytes, dst-bytes, land, hot, num-failed-login, logged-in, num-compromised, Root-shell, su-attempted, num-root, num-file-creation, num-shells, num-access-files, is-host-login, is-guest-login*. So, when these features are used with Naïve Bayes, we can obtain results which have been represented in Table 6:

Table 6. Analysis InfoGain + Naïve Bayes on NSL-KDD

Attacks	InfoGain + Naïve Bayes	Naïve Bayes with 40 features
U2R	57%	14%
R2L	19%	17.5%
DoS	73.4%	70%
Probe	74%	86.5%

In these experiments, various tests with using different feature selection methods to select the best features are accomplished. However, when the evaluation is done based on “SVM”, “Decision Trees”, “Neural Networks” for the detection of Probe and DoS Attacks, have no good results are obtained.

7 Conclusion

Regardless of KDDCUP99 data sets defects, such as data redundancy and duplicated records, among the mentioned techniques based on McHugh and A.A.Ghorbani et al. reports in [6] and [7] respectively and also according to investigation conducted on KDDCUP99, it is concluded that decision trees and SVM work outperform other methods for detecting the normality of input packet. Also, Neural Networks have a better performance than other methods for detection of *DoS* attacks. Similarly, for the *Probe* attacks, “Decision Trees” are much better than the other methods. Meanwhile, “Naïve Bayes” is also the most effective method for detecting *U2R* and *R2L* attacks. The result of conducted evaluations on NSL-KDD data set shows that Feature Selection techniques in NSL-KDD data set cause problems at detection of *probe* attacks. It can be concluded that among mentioned techniques and investigations that have been conducted for the detection of normality of input packet and also detection of *DoS* attacks, decision trees report a better result than other techniques. For *Probe* attacks, "Naïve Bayes" technique is better than the others and for *U2R* and *R2L* attacks; “InfoGain” and "Naïve Bayes" techniques have better results. For detecting *DoS*, *Probe*, normality input packets all the features except feature *num_outbound_cmds* should be used. This summary is shown in Fig. 2:

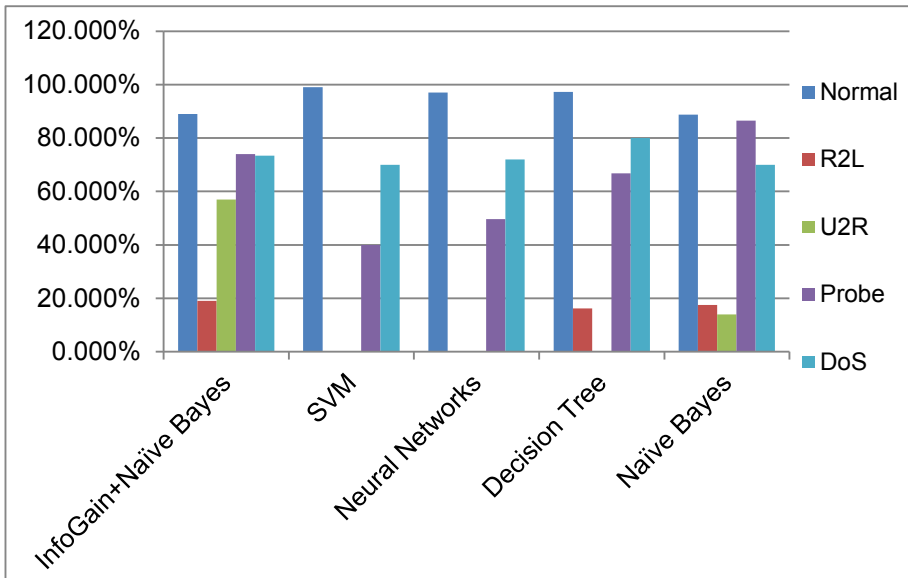


Fig. 2. The comparison some machine learning techniques in category of attacks

References

1. Lazarevic, A., Ertoz, L., Ozgur, A., Srivastava, J., Kumar, V.: A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of Third SIAM Conference on Data Mining (2003)
2. Sabhnani, M., Serpen, G.: Application of machine learning algorithms to KDDCUP99 intrusion detection dataset within misuse detection context. In: International Conference on Machine Learning Models, Technologies and Applications Proceedings, pp. 209–215 (2004)
3. KDDCUP99 Dataset. <https://kdd.ics.uci.edu/databases/kddcup99>
4. Stolfo, S.J., Fan, W., Lee, W., Prodromidis, A., Chan, P.K.: Cost based modeling for fraud and intrusion detection: Results from the jam project, disceX, vol. 02, p. 1130 (2000)
5. Shyu, S., Chen, K., Sarinnapakorn, K., Chang, L.: A novel anomaly detection scheme based on principal component classifier. In: Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, in conjunction with the Third IEEE International Conference on Data Mining (ICDM 2003), pp. 172–179 (2003)
6. Tavallae, M., Bagheri, E., Lu, W., Ghorbani, A.: A detailed analysis of the KDDCUP99 data set. In: Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA) (2009)
7. McHugh, J.: Testing intrusion detection systems: a critique of the 1998 and 1999 Darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Trans. Inf. Syst. Secur.* **3**(4), 262–294 (2000)
8. Nsl-kdd data set for network-based intrusion detection systems. <https://nsl.cs.unb.ca/NSL-KDD>
9. Adetunmbi Adebayo, O., Shi, Z., Shi, Z., Adewale, O.S.: Network anomalous intrusion detection using fuzzy-bayes. In: Shi, Z., Shimohara, K., Feng, D. (eds.) IIP 2006. IIFIP, vol. 228, pp. 525–530. Springer, Boston, MA (2006). https://doi.org/10.1007/978-0-387-44641-7_56

10. Amor, N.B., Benferhat, S., Elouedi, Z.: Naïve Bayesian Networks in Intrusion Detection Systems. In: 14th European Conference on Machine Learning (Dubrovnik) (2003)
11. Pachghare, V.K., Kulkarni, P.: pattern based network security using decision trees and support vector machine, Department of Computer Engineering And Information Technology College of Engineering, Pune, India. IEEE (2011)
12. Werbos, P.J.: Beyond regression. New tools for prediction and analysis in the behavioral sciences, Ph.D. thesis, Harvard University (1974)
13. Rajesh, R., Sheen, S.: Network Intrusion Detection using Feature Selection and Decision tree classifier (2008)
14. Teng, S., Du, H., Wu, N., Zhang, W., Su, J.: A cooperative network intrusion detection based on fuzzy SVMs. *J. Networks* **5**(4), 475, Academy Publisher (2010)
15. Chen, R.-C., Cheng, K.-F., Hsieh, C.F.: Using rough set and support vector machine for network intrusion detection. *Int. J. Network Secur. Appl. (IJNSA)* **1**(1) (2009)
16. Sabhnani, M., Serpen, G.: Application of machine learning algorithms to KDDCUP99 intrusion detection dataset within misuse detection context. In: International Conference on Machine Learning, Models, Technologies and Applications Proceedings, pp. 209–215 (2004)
17. Alfantookh, A.A.: DoS Attacks Intelligent Detection using Neural Networks, Department of Computer Science, College of Computer and Information Sciences King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia (2005)
18. Mukhopadhyay, I., Chakraborty, M., Chakrabarti, S., Chatterjee, T.: Back propagation neural network approach to intrusion detection system. In: International Conference on Recent Trends in Information Systems, Department of Information Technology, Institute of Engineering and Management (2011)
19. Kayacık, H.G., Zincir-Heywood, A.N., Heywood, M.I.: Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDDCUP 99 Intrusion Detection Datasets (2005)