ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Performance Evaluation for Novel GA based Intrusion Detection System

[1]S.Vijayarangam and [2]Dr.A.Rajesh

[1]*Research Scholar, Department of Computer Science and Engineering, St.Peter's University, Avadi, Chennai.*
[2]*Professor & Head, Department of Computer Science and Engineering, C.Abdul Hakeem College of Engineering and Technology, Melvisharam, Tamilnadu.*

ARTICLE INFO

Article history:
Received 12 July 2015
Accepted 28 August 2015
Available online 15 September 2015

Keywords:
Intrusion Detection, Genetic Algorithm, KDD99, Training Accuracy.

ABSTRACT

The need for an efficient Intrusion Detection System arises to preserve data integrity and system reliability for computer network and security. To compete with the commercial intrusion detection techniques, immense care should be taken on reducing the computational overhead. Under this decisive factor we propose a novel Genetic Algorithm for intrusion detection to effectively detect various types of computer network intrusions. The conventional technique utilizes some existing feature extraction techniques to reduce the amount of data in the process. Here we propose a novel feature extraction and classification algorithm. First the binary host event data is first processed into ASCII information; this is reviewed in terms of service and duration and summarized into host session records. The uniqueness to reduce the computational overhead is maintained by applying our data mining algorithm after the preprocessing stage. Our data mining algorithm is applied to host session records for estimating the recurrent patterns in the network. These recurrent patterns are further analyzed to build supplementary features for the host session records. Then the classification rule is applied to learn the detection model. KDD99 dataset has been used for benchmark test verification. To evaluate our system, standard metric like training accuracy, testing accuracy, detection rate and false positive rate will be estimated for fair comparison. And it is found that, the proposed system delivers better accuracy under 22 out of 41 features testing case.

To Cite This Article: S.Vijayarangam, Dr.A.Rajesh., Performance Evaluation for Novel GA based Intrusion Detection System. ***Aust. J. Basic & Appl. Sci.,*** *9(27): 778-784, 2015*

INTRODUCTION

The internet security faces challenges of fast changing trends of attacking the internet resources. There is an inability of conventional techniques to protect the internet resources from a variety of attacks. An efficient technique is necessary for securing valuable internet resources from attacks. For the security of the system, some conventional protection techniques like firewalls, authentication system and encryption techniques are used as the first line of defense. This first line of defense prevents from some attacks whereas some bypass them. These kinds of attacks must be detected soon, so that we may able to minimize the damage. Several techniques are being employed from different disciplines for the accurate intrusion detection system. To evaluate the capability of intrusion detection system, two key indicators called Detection Rate (DR) and False Positive Rate are needed (Patcha, A. and J.M. Park, 2007).

In the beginning days, the researchers focused on rule based and statistical intrusion detection systems. But, the results become unsatisfactory, with large datasets. After that, artificial intelligence based techniques have been introduced, and shows certain improvement in detecting the intrusions (Wu, S.X. and W. Banzhaf, 2010). Most of the existing techniques make great efforts to achieve a single solution (Engen, V., 2010). There was no single technique to detect all kinds of attacks to a certain level of detection accuracy and false alarm rate (Re, M. and G. Valentini, 2010), also they are not capable of modeling correct hypothesis space of the problem (Dietterich, T and G. Bakiri, 1994). Some of the techniques are unstable; some could not process the larger size like high dimensional data (Chandola, V., et al., 2009). The Genetic Algorithm is also implemented to carry out the intrusion detection system. GA performed effectively and identified various types of intrusions (Gomez, J. and D. Dasgupta, 2002), which is applied to achieve a set of classification rules and for the selection of appropriate features of the chromosome. This paper proposes a novel Genetic Algorithm which utilizes

Corresponding Author: S.Vijayarangam, Department of Computer Science and Engineering, St.Peter's University, Avadi, Chennai.
E-mail: rndvijayarangam@gmail.com

some existing feature extraction techniques to reduce the amount of data in the process.

2. Literature Review:

Previous studies in the intrusion detection field have come across many techniques to generate effective ensembles. Ensemble techniques or classifiers have been applied to overcome the limitations of a single classifier (Dietterich, T., 2000). Roli (2001) proposed a multi classifier based system of neural networks. The various neural networks were used different features of KDD cup 99 datasets. This paper concluded that a multi strategy combination technique like belief function outperforms other representative techniques. In (Gomez, J., et al., 2002) J. Gomez et al. proposed a linear representation scheme for evolving fuzzy rules using the concept of complete binary tree structure. Genetic Algorithm is used to generate genetic operators for producing useful and minimal structural modification to the fuzzy expression tree represented by chromosomes. The biggest limitation of the proposed approach was that the training was time consuming. Middlemiss et al (2003) have used GA for weighted feature extraction with specific application to intrusion detection. A k-nearest neighbor classifier was used for the fitness function of GA as well as to evaluate the performance of the new weighted feature set. These weighted features are used to scale the input variables provided to the classifier system.

Xiao et al. (2005) detect anomalous network behavior based on information theory using Genetic Algorithm. Some network features can be identified with network attacks based on mutual information between network features and type of intrusions and then using these features a linear structure rule and also a Genetic Algorithm is derived. The approach of using mutual information and resulting linear rule seems very effective because of the reduced complexity and higher detection rate. The main problem is, it considered only the discrete features. Hu and Damper (2008) proposed an AdaBoosting ensemble method that uses different features to generate a diverse set of classifiers. The proposed method reported improved performance but it suffers from the limitation of incremental learning. The only drawback is, it requires continuous retraining for a changing environment.

Wang et al. (2010) proposed an approach based on NN and fuzzy clustering, which helps to generate homogeneous training subsets from heterogeneous training datasets which are further used to train NN models. This paper reported performance improvement in terms of detection precision and stability. In (Mohammad Sazzadul Hoque, et al., 2012) Hoque, Mukit and Bikas presented an implementation of Intrusion Detection System by applying the theory of genetic algorithm to efficiently detect various types of network intrusive activities. To apply and measure the efficiency of their system they used the standard KDD 99 intrusion detection benchmark dataset and obtained realistic detection rate. This paper measured the fitness of a chromosome using the standard deviation equation with respect to distance. But the detection rate was poor and they failed to reduce the false positive rate in Intrusion Detection System.

3. Understanding KDDCUP99 Data:

KDD99 is built based on the data captured in DARPA'98 IDS evaluation program. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcp dump data of seven weeks of network traffic that can be processed into about 5 million connection records and each about 100 bytes.

And the remaining two weeks of test data have around 2 million connection records. The KDD training dataset was approximately consists of 4,900,000 single connection vectors each of which contains 41 features (34 *features* are numeric and 7 *features* are symbolic) and is labeled as either normal or an attack, with exactly one specific attack type (Mahbod, T., et al., 2009).

KDD has evolved from interaction and cooperation among such different fields like pattern recognition, database, statistics, AI, and knowledge acquisition for intelligent systems. Discovering a high level knowledge from lower levels of relatively raw data, or to discover a higher level of abstraction and interpretation than those previously known, is the main idea in KDD. KDD applies machine learning and pattern recognition techniques to extract patterns implicit in a database. The new wave of KDD addresses the overall process of discovering useful knowledge from data while data mining, pattern recognition and other such techniques address only a particular step. KDD seeks incrementally to understand, to apply and to adapt these patterns to future cases or data sets (Al-mamory, Safaa O., and Firas S. Jassim). Statistical methods and algorithms offer precise methods for quantifying inherent inferential uncertainties. The KDD systems embed with the particular statistical procedure for and modeling data and handling noise within an overall knowledge discovery framework. KDD approaches and methods are focused on model extraction or discovery, instead of the parameter estimation of previously hypothesized model. They gave their best operational performance in the context of large sets with rich data structures. In some kind of large data sets, interpretations may already exist. By shifting the window of concern to another aspect of the database, we may have chance to get some new pattern for another purpose.

Table 1: The 41 Features

S.NO	Feature Name	S.NO	Feature Name
1	Duration	22	Is_guest_login
2	Protocol type	23	Count
3	Service	24	Serror_rate
4	Src_byte	25	Rerror_rate
5	Dst_byte	26	Same_srv_rate
6	Flag	27	Diff_srv_rate
7	Land	28	Srv_count
8	Wrong_fragment	29	Srv_serror_rate
9	Urgent	30	Srv_rerror_rate
10	Hot	31	Srv_diff_host_rate
11	Num_failed_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_count
13	Num_compromised	34	Dst_host_same_srv_count
14	Root_shell	35	Dst_host_diff_srv_count
15	Su_attempted	36	Dst_host_same_src_port_rate
16	Num_root	37	Dst_host_srv_diff_host_rate
17	Num_file_creations	38	Dst_host_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_shells	40	Dst_host_rerror_rate
20	Num_outbound_cmds	41	Dst_host_srv_rerror_rate
21	Is_hot_login		

4. Data Mining Based IDS:

The statistical detection technique is based on the behavior of a user's or a system's action, which significantly deviates from the normal behavior. This technique focuses on normal behavior patterns. An activity will be treated as an intrusion and will give a false positive result, when a new kind of normal behavior pattern is not updated in the database. This technique also struggles from several limitations, even though it can detect unknown patterns of intrusion. There is a general problem of this detection technique is, the normal behavior is modeled on the basis of the data collected over a period of normal operations, so if any attack occurs during this period, it will be taken the attack as a normal activity. A user's behavior may change when his/her projection changes and it changes over time. So, this may lead an intruder to gradually train and he/she may trick it. Because of some technical reasons, the current detection approaches usually suffer from a high false-alarm rate. A data mining based Intrusion detection is seriously more complex than a traditional system. The data mining system needs large sets of training data. It is very difficult to manage both historical and training data. After the new data has been analyzed, models need to be updated. It is really impossible to retrain all the available data for updating the models, because retraining can take some weeks or even months (Helali, Rasha G. Mohammed, 2010). The data mining based intrusion detection is very difficult to deploy because they require a large set of labeled clean data.

5. Proposed Work:

The problem of intrusion detection is not just identifying the attacks, but also to know the type of the connection. A genetic algorithm is a very good way to find an efficient solution to the problem. This process usually begins with the chromosomes which are randomly generated and represent all the possible solutions of the problem. Different positions are encoded as bits, numbers or characters from each chromosome. The goodness of each chromosome is calculated according to the desired solution using evaluation function, which is known as fitness function. Selection, crossover and mutation are the genetic operators that are sequentially applied to each individual with some certain probabilities (Omprakash Chandrakar, et al., 2014). Along with these IDS using GA approach a strong classification algorithm is proposed. This classification algorithm distinguishes important alerts from redundant one. It is a data mining technique used to map data instances in one of the various predefined categories. It can be used to detect individual attacks but it has a high rate of false alarm. It employs frequent item set mining for detecting patterns that describe frequently occurring redundant alerts. The classification algorithm has been then applied to audit data collected which then learns to classify new audit data as normal or abnormal data. The main aim of this algorithm is to reduce the false positive rate. The false positive rate is basically known as the percentage of a true thing which is wrongly recognized as an attack. This can be calculated as,

$$False\ Positive\ Rate = \frac{\#False\ Positive}{\#True\ Negative + \#False\ Positive} * 100$$

If the false positive rate is increased the alert system has to be called for the positive connections too. This algorithm will first get the event data of the host in each and every session and convert that information into session records, which describe strong associations between alert attribute values for each host ID. Such session records consist of host-bound audit sources such as operating system audit trails, system logs, or application logs, IP addresses, source, duration, timestamp and flags. This will help to understand the behavior of the host so that

classification algorithm can alert the data mining algorithm to save the particular data, if the source behavior crosses a certain limit of the classification rule. A classification algorithm classifies the intrusion into some matrices that are, false positive, false negative, and true positive and true negative. Classification algorithm will form rules under these matrices, which is called as classification rules. This classification rule will apply to the data mining algorithm. So the whole mining system will be under a smart alert. We have experimented this proposed system to test its effectiveness on reducing false positive rate and got results as shown in section 6.

The main advantage of using rules with knowledge base is that it helps to perform effective decision making on intrusions. Learning subset and testing subset were arbitrable selected from two subsets of the full data. In the next step, the features which have symbolic form were transformed into numerical values by allocating a unique number for each feature. This method is deployed in both learning and testing subset. To maintain the effectiveness, it is recommended to scale the input features to fall within a specific range. Then the equation (2) is developed to normalize the features in the learning and testing data set.

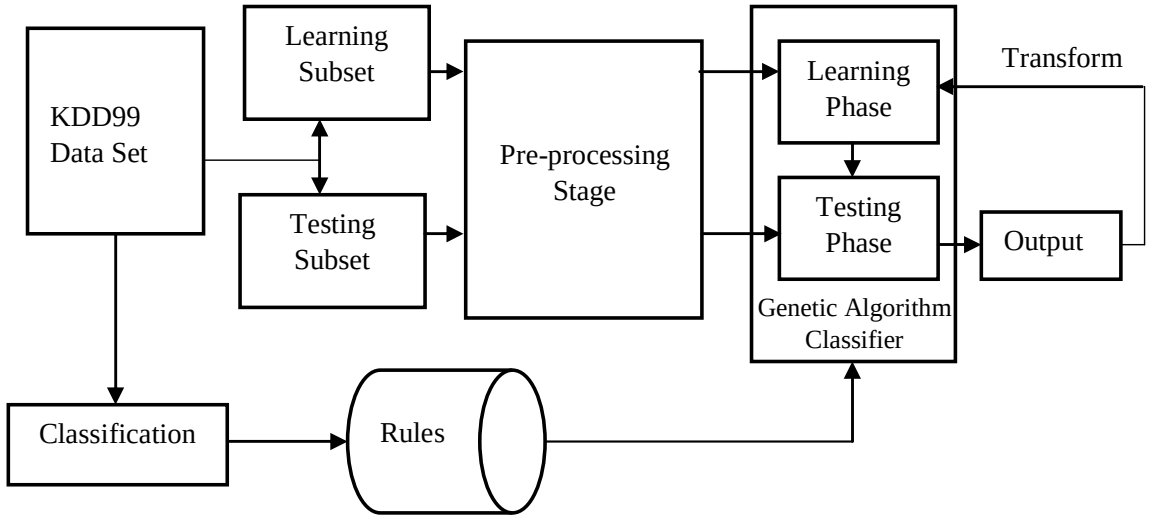

Fig. 1: Proposed Model

$$\text{Rational Difference} = (DVal - Dval_{max}) / (Dval_{max} - Dval_{min}) \quad (1)$$

$$\text{Normalized Output } (D_{norm}) = 2 * \text{Rational difference} \quad (2)$$

Where,

$DVal_{min}$, $DVal_{max}$ are the maximum and minim value of the original inputs

$DVal_{norm}$ is the normalized output.

The normalized output will fall in the range [-1, +1]. The training and learning dataset will have $(f,R_t,I_t)$and (f,R_l,I_l) respectively. Where *f :*features , R_t : Training Records , I_t : Training Intrusion type , R_l : Learning Records, I_l : Learning Intrusion type. Classification rules are prepared from the available data using the GA in an offline environment. In the real time environment, these rules are used to classify the inward network connections. With reference from (Chittur, A., 2006), we have deployed the fitness value function to estimate the fitness of each rule.

$$\text{Fitness} = \frac{\text{Correctly detected attacks}}{Totalattacks\ in\ the\ training\ data\ set} - \frac{false\ attacks}{Total\ normal\ connections\ in\ the\ training\ dataset} \quad (3)$$

The range of the fitness values will be [-1,+1]. For a high fitness value (+1), high detection rate and low rate of false-positive is needed. Inversely, low detection rate and high rate of false-positive result in low fitness value (-1). The 41 features of the KDD cup data set are loaded in each network connection which will be either normal or attack.

RESULTS AND DISCUSSION

The experimental results are tabulated in Table 2. The proposed method is compared with the other conventional techniques such as BayesNet, J48 and GAIDS in the weka software for case 1, 2 &3. In particular we focus on training accuracy, testing accuracy, detection rate and false positive rate.

Table 2: Evaluation Results

		BayesNet	J48	GAIDS	Proposed Method
Case1	Training Accuracy	99.978	99.983	99.9695	99.981
	Testing Accuracy	99.9839	99.58	99.971	99.983
	Detection Rates	99.98	99.98	99.87	99.94
	False Positive Rates	0.021	0.006	0.003	0.0017
Case2	Training Accuracy	99.891	99.9914	99.9497	99.9984
	Testing Accuracy	99.9172	99.891	99.89	99.987
	Detection Rates	99.94	99.982	99.85	99.971
	False Positive Rates	0.189	0	0.024	0.0015
Case3	Training Accuracy	99.7586	99.9977	99.582	99.994
	Testing Accuracy	99.6744	99.987	99.67	99.986
	Detection Rates	99.90	99.991	99.82	99.904
	False Positive Rates	0.293	0	0.5	0.014

During the first case of our experiment GAIDS and our proposed method has shown better performance in both training and testing accuracy. The detection rate of bayes, J48 and our system were comparatively better as shown in the above graph which is named as Fig. 2.

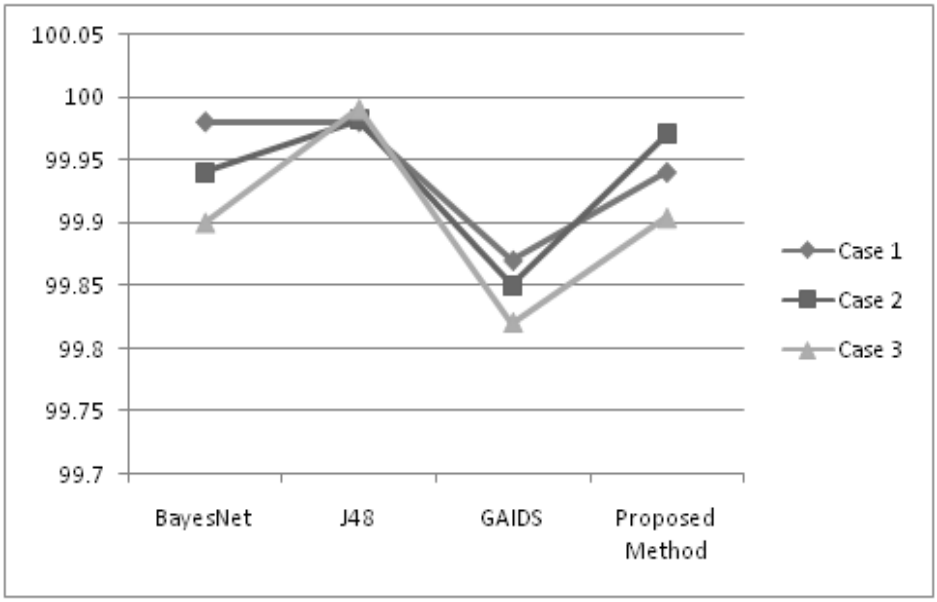

Fig. 2: Comparison of Detection Rate

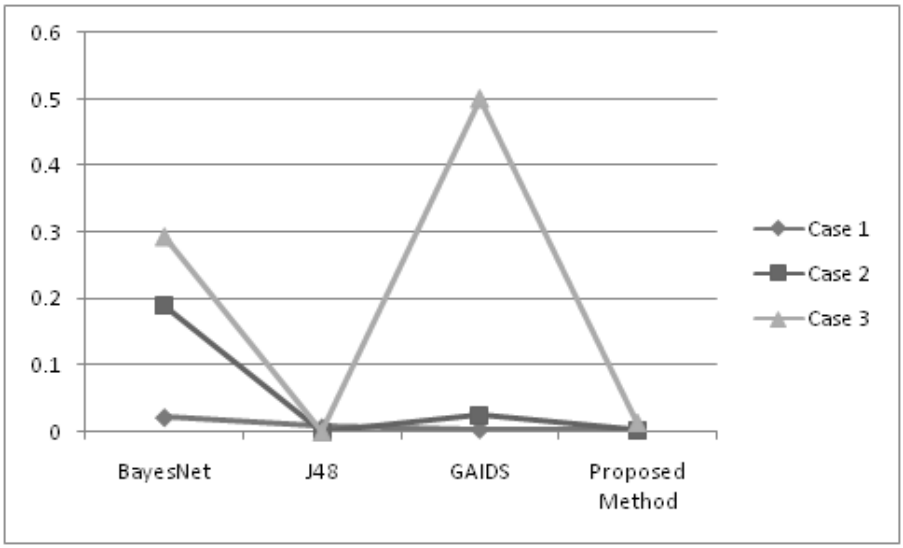

Fig. 3: Comparison of False Positive Rate

J48 and our method performed really well in all the cases of our experiments and those methods gave a very small range of false positive rate, which is graphically represented in Fig. 3.

Conclusion:

In this paper, we have proposed a novel technique of deploying genetic algorithm to detect the network intrusion. We have investigated new techniques for intrusion detection and evaluated their performance in terms of training accuracy, testing accuracy, detection rates and false positive rates. We have used the feature extraction and classification algorithm in the KDD99 Cup data set to evaluate the proposed GA based intrusion detection system. Three different tests cases i.e. 18 out of 41 features, 22 out of 41 features and 31 out of 41 were simulated and our proposed system has a higher training accuracy of 99.9984% in case 3, than the conventional GA based intrusion detection system. Our proposed system is capable of achieving a lower false positive rate of 0.0017%, 0.015% and 0.14% in the three testing cases. From the empirical results, it is proved that the proposed GA based intrusion detection system has the best training and testing accuracy under 22 out of 41 features. However in the future, our research work will be focused on developing more prominent intrusion detection system to achieve 100% accuracy.

REFERENCES

Al-mamory, Safaa O., and Firas S. Jassim, "Evaluation of Different Data Mining Algorithms with KDD CUP 99 Data Set".

Chandola, V., A. Banerjee and V. Kumar, 2009. "Anomaly detection: a survey", ACM Computing Surveys, article, 15, 41(3).

Chittur, A., 2006. "Model Generation for an Intrusion Detection System Using Genetic Algorithms".

Dietterich, T and G. Bakiri, 1994. "Error-correcting output codes: a general method for improving multiclass inductive learning programs", in Proceedings of the of Santa fe Institute Studies in the Sciences of Complexity, Citeseer, pp: 395-395.

Dietterich, T., 2000. "Ensemble methods in machine learning", in Proceedings of Workshop on Multiple Classifier Systems, pp: 1-15.

Engen, V., 2010. "Machine learning for network based intrusion detection: an investigation into discrepancies in findings with the KDD cup'99 data set and multi-objective evolution of neural network classifier ensembles from imbalanced data [Ph.D. thesis]", Bournemouth University.

Giacinto, G. and F. Roli, 2001. "Approach to the automatic design of multiple classifier systems", Pattern Recognition Letters, 22(1): 25-33.

Gomez, J. and D. Dasgupta, 2002. "Evolving Fuzzy Classifiers for Intrusion Detection", IEEE Proceedings of the IEEE Workshop on Information Assurance, United States Military Academy, West Point, NY.

Gomez, J., D. Dasgupta, D. Nasaroui and F. Gonzalez, 2002. "Complete expression Trees for evolving Fuzzy classifiers system with Genetic Algorithms and Applications to Network Intrusion Detection", pp: 469-474.

Helali, Rasha G. Mohammed, 2010. "Data mining based network intrusion detection system: A survey. Novel Algorithms and Techniques in Telecommunications and Networking". Springer Netherlands, pp: 501-505.

Hu, R. and R.I. Damper, 2008. "A 'No Panacea Theorem' for classifier combination", Pattern Recognition, 41(8): 2665-2673.

Mahbod, T., E. Bagheri, Wei Lu and A.A. Ghorbani, 2009. A Detailed Analysis of the KDD CUP 99 Data Set", pp: 2.

Melanie Middlemiss, Grant Dick, 2003. "Weighted Feature Extraction Using a Genetic Algorithm for Intrusion Detection", 2003 Congress on Evolutionary Computation (cec-03), pp: 1669-1675.

Mohammad Sazzadul Hoque, Md. Abdul Mukit & Md. Abu Naser Bikas, 2012. "An Implementation of Intrusion Detection System using Genetic Algorithm", International Journal of Network Security and Its Applications (IJNSA), 4(2): 109-120.

Omprakash Chandrakar, Rekha Singh, Dr. Lal Bihari Barik, 2014. "Application of Genetic Algorithm in Intrusion Detection System", International Institute for Science, Technology and Education, 4(1): ISSN. 2224-5774.

Patcha, A. and J.M. Park, 2007. "An overview of anomaly detection techniques: existing solutions and latest technological trends', Computer Networks, 51: 3448-3470.

Re, M. and G. Valentini, 2010. "Integration of heterogeneous data sources for gene function prediction using decision templates and ensembles of learning machines", Neurocomputing, 73(7-9): 1533-1537.

Wang, G., J. Hao, J. Mab and L. Huang, 2010. "A new approach to intrusion detection using

artificial neural networks and fuzzy clustering", Expert Systems with Applications, 37(9): 6225-6232.

Wu, S.X. and W. Banzhaf, 2010. "The use of computational intelligence in intrusion detection systems: a review", Applied Soft Computing Journal, 10(1): 1-35.

Xia, T., G. Qu, S. Hariri, M. Yousif, 2005. "An Efficient Network Intrusion Detection Method Based on Information Theory and Genetic Algorithm", Proceedings of the 24th IEEE International Performance Computing and Communications Conference (IPCCC, 05), Phoenix, AZ, USA.