Asian Journal of Information Technology 10 (4): 136-141, 2011
ISSN: 1682-3915

Design and Implementation of a Data Mining-Based Network Intrusion Detection Scheme

Rasha G. Mohammed and Awad M. Awadelkarim
College of Computer Science and Information Technology,
Sudan University of Science and Technology, P.O. Box 407, Khartoum, Sudan

Abstract: A significant security problem for networked systems is hostile trespass by users or software. Intruder is one of the most publicized threats to security. In actual fact, most of the current systems are weak at detecting novel attacks without generating false alarms. Intrusion Detection Systems (IDSs) are increasingly a key part of systems defense. Various approaches to intrusion detection are currently being used which are relatively ineffective. Likewise, data mining plays a driving role in data analysis. This study addresses this issue and proposes a data mining-based intrusion detection system. The data mining techniques being investigated include decision tree (C5.0 algorithm) and distance based clustering (Tow-steps algorithm). The proposed hybrid system combines anomaly and misuse detection. Experiments are performed on both real network data for Sudan University of Science and Technology (SUST) network and Defense Advanced Research Projects Agency (DARPA) dataset which is considered as the most famous available off-line intrusion detection evaluation dataset. The obtained results confirm that data mining is capable of discovering attacks with acceptable level of false alarms.

Key words: Anomaly detection, data mining, misuse detection, intrusion detection system, network intrusion detection system, Sudan

INTRODUCTION

The amount of data being collected in databases today far exceeds the ability to direct and analyze data without the use of automated analysis techniques such as data mining. Data mining is defined as the nontrivial extraction of implicit, previously unknown and potentially useful information from large data sets or databases (Dunham, 2003).

The ongoing rapid development in data mining has made available a wide variety of algorithms in such area which drawn from the fields of statistics, pattern recognition, machine learning and databases. Numerous techniques are used to implement such algorithms as the following:

- Classification which maps a data item into one of several predefined categories. Classification algorithms normally yield classifiers that have ability to classify new data in the future such as in the form of decision trees or rules
- Clustering which maps data items into groups according to similarity or distance between them. The best way of finding out the deviation from normal use of a network (anomaly detection) is to use clustering technique
- Link analysis determines relations between fields in the database
- Sequence analysis models sequential patterns. These event patterns are important when creating behavior profile of a user or program (Zhu *et al*., 2001)

Data mining mechanisms such as rule induction, neural networks, genetic algorithms, fuzzy logic and rough sets are used for classification and pattern recognition in many industries such as business intelligence organizations and financial analysts. Also, they are increasingly used in the sciences to extract information from the enormous datasets generated by modern experimental and observational methods. Data mining has been extensively used in discriminating normal from abnormal behavior in a variety of contexts (Zhu *et al*., 2001). In recent years data mining techniques have been successfully used in the context of network intrusion detection (Shyu and Sainani, 2009; Shanmugam and Idris, 2009; Prasad *et al*., 2008).

Intrusion detection includes identifying a set of malicious actions that compromise the integrity, confidentiality and availability of information resources. The traditional methods for intrusion detection are based on extensive knowledge of signatures of previously

Corresponding Author: Rasha G. Mohammed, College of Computer Science and Information Technology, Sudan University of Science and Technology, P.O. Box 407, Khartoum, Sudan

known attacks. Monitored events are matched against the signatures to detect intrusions. These methods extract features from various audit streams and detect intrusions by comparing the feature values to a set of attack signatures provided by human experts. The signature database has to be manually revised for each new type of intrusion that is discovered (Dokas and Ertoz, 2002).

Generally, a significant limitation of such methods is that they cannot detect emerging cyber threats which do not have signatures or labeled data corresponding to them. In addition, even if a new attack is discovered and its signature developed often there is an estimated latency in its deployment across networks (Dokas *et al*., 2002). Moreover, the current (conventional) techniques have several limitations such as producing loads of false alarms and they need extensive training data for the associated algorithms. These limitations have led to an increasing interest in intrusion detection techniques based on data mining in place of the conventional methods in order to reveal attacks efficiently.

Thus, this study proposes a framework that uses data mining for building Network Intrusion Detection System (NIDS) for Sudan University of Science and Technology (SUST) Network.

RELATED WORK

Historically, the intrusion detection technology dates back to 1980 and becomes a well-established research area at the ends of 1980's (Eid, 2004). Sights moved for using data mining in context of NIDS in the late of 1990's. Researchers quickly recognized the need for exisitance of standardized datasets to train IDS tool. First widely cited datasets for the information exploration shootout have been concederd in which unfortunately is no longer available (Lee and Stolfo, 1998). In the most famous avilable datasets Defense Advanced Research Projects Agency (DARPA) have been mentioned (Brugger, 2004). It was made available to researchers in 1998 as the DARPA off-line intrusion detection evaluation dataset. Many researchers use DARPA/KDD dataset which appears to be the most useful dataset that can be used without any further processing (Giacinto and Roli, 2002; Shanmugam and Idris, 2009).

A great deal has been done by Lee whom analysing DARPA dataset and identifing 41 features which can be used in a data mining based NIDS. A copy have been provided for the 1999 Knowledge Discovery and Data Mining cup 1999 KDD Cup held at the 5th association for computing machinery ACM international conference on Knowledge Discovery and Data Mining (Fig. 1).

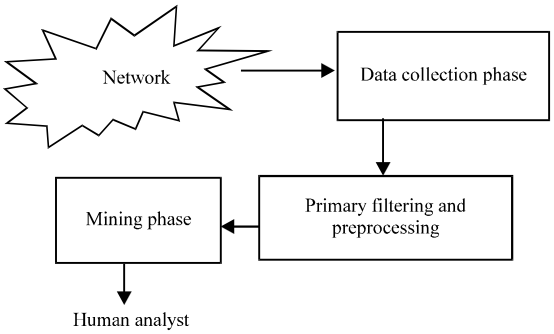


Fig. 1: Deduced architecture of data mining NIDS

Several studies have compared the performance of various data mining methods. An attempt was made in (Dokas and Ertoz, 2002) to develop a model which focuses on the prediction of rare classes to identify known intrusions and their variations anomaly-outlier-detection. The study started by a comparative study to identify several anomaly and misuse detection schemes to determine the best scheme to be used. The results proofed the success of some algorithms over others such as Synthetic Minority Over-Sampling Technique (SMOTE) classification algorithm for misuse and Nearest Neighbor NN, Density Based Local Outliers LOF for anomaly detection. In a systematic method for intrusion detection is presented by using data mining techniques (Lee *et al*., 1998). Their experiments emphasize that accuracy of detection model depends on sufficient training data and feature set. In the basic association rules and Frequent Episodes algorithms are extended to accommodate the special requirements in analyzing audit data for both misuse and anomaly detection (Lee *et al*., 1999). The study discover that using of multiple classifiers is best than using single one for detecting attacks.

Minnesota Intrusion Detection System (MINDS) touched in provides a good example of combining signature based tool with data mining techniques (Ertoz *et al*., 2004). Signature based tool (Snort) are used for misuse detection and data mining for anomaly detection. The framework contains several integral parts such as filtering/preprocessing including extraction of new features what is known as time based features and connection based features, Known attack detection module (using Snort), Anomaly detection algorithms (using scoring algorithm that identify the most anomalies connections) and summarization of attacks using association pattern analysis. It enjoys great operational success in detecting brand new attacks that signature-based systems could not have found. A great contribution is offered in though building a simple framework to get started in data mining based NIDS.

In data mining-based NIDS in real time are discussed by using algorithms adopted from Apriori which are presented by Peng and Zuo (2006) and Jiawei and Kamber (2001). In context of integrating fuzzy logic in NIDS different attempts were made. In an attempt was made by Idris and Shanmugam (2006). They proposed a dynamic intelligent intrusion detection system model mixed between anomaly and misuse detection techniques and fuzzy logic. Their initial experiments show promising and encouraging results. A similar idea has been pursued by Prasad *et al*. (2008). Genetic algorithms based on fuzzy logic was used to produce better results. Based on previous studies concurs, researchers can conclude that although there are differences in architectural details of used frameworks which are constructed for different environments; they agree in general steps that can be abbreviated as follows:

- Data collection phase
- Data filtering and preprocessing
- Mining phase

THE PROPOSED FRAMEWORK

This research describes the proposed framework that has been adopted thought this study. The framework investigated here benefits from the frameworks being conducted as a part of MINDS (Minnesota Intrusion Detection System) project and Massachusetts Institute of Technology Research and Engineering (MITER) network. Minnesota intrusion detection system which proves its success during experiments done at Minnesota University are taken as beginning step of building the proposed framework suit for Sudan University of Science and Technology SUST network. The proposed framework combines misuse detection phase which is used for filtering previously known attacks by using data mining instead of signature based tool used in MINDS and Anomaly-detection phase. Figure 2 shows the proposed framework applied to SUST network.

The proposed framework phases

Data collection phase: Captureing is performed using a special sniffer developed using C language. The sniffer is used to capture all packets and store its header only in a MySQL database. The captuerd featuers for every packet include source and distnation IP, soucre/destination port, protocol, number of bytes, service type and flag.

Data filtering phase: Captured data are filtered in order to remove network traffic non relevant for analysis. This study concentrates on Transmission Control Protocol (TCP), User Datagram Protocol (UDP) and Internet Control Message Protocol (ICMP) packets because the majority of connections in SUST network fall within these protocols.

Feature extraction phase: At this step new features are extracted to prepare for both mining step misuse and anomaly detection. Extracted features are time and connection-based features. Since network connections are already captured during specified time period and processed offline there is no need to use both time and

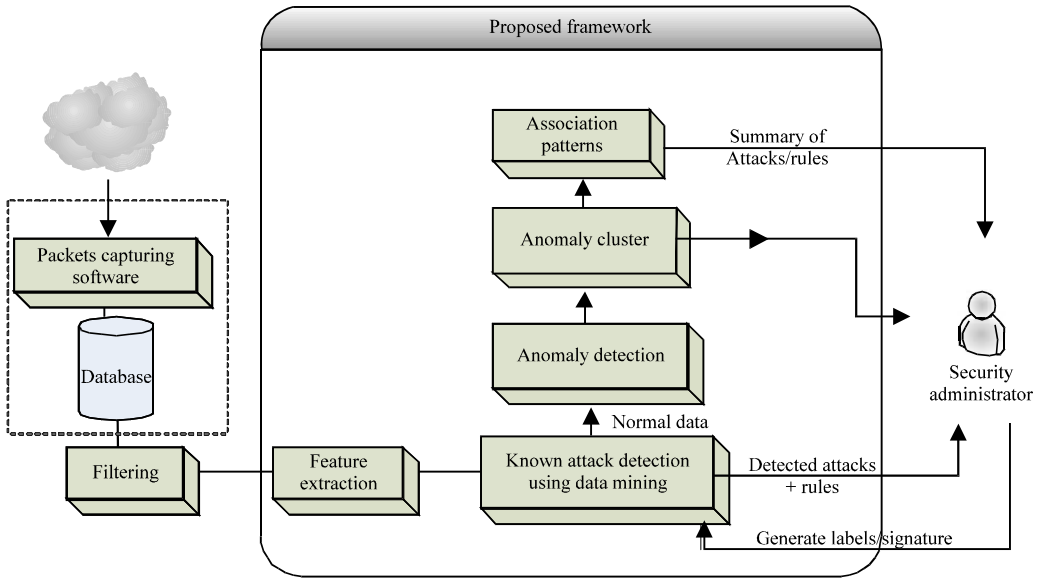


Fig. 2: Proposed framework

connection based feature. Instead, only connection-based features are used. Certain rules have been suggested by the network and security administrator to identify some types of attacks such as Denial of Services DoS and Worms. DoS rules are assumed based on the nature and traffic of the network. Following is a sample of these rules:

- Rule 1: any packets going to port 21 with service FTP represent attacks against an FTP server (prevent any FTP connection)
- Rule 2: connections that flow from a same source IP to a same destination IP in 90 min with a total ≥30 with service http/ping specify as intrusive connections (DoS)
- Rule 3: any telnet connection represented as intrusions (prevent telnet connections)
- Rule 4: any connection flows from any source port to destination port 445 (Worm)
- Rule 5: connections with SYN flag ≥30 represents intrusive connection (SYN flood). Rules are used initially to label each network connection as normal or intrusion

Data mining phase

Known attacks detection model: The known attacks detection model classifying network data as normal or intrusion based on labeled training data. Decision tree classification algorithm C5.0 is used to perform classification process. C5.0 algorithm has boosting capability which means classifying data internally into multiple sets and using combination of classifiers to increase the classification accuracy. The out put of this phase is two sets of connections normal and detected attacks.

Anomaly detection model: Clustering technique is used to discover new attacks not detected as intrusions in the previous step. The Tow-step clustering algorithm is used. It has the ability to group data into a number of clusters automatically based on distance criteria. The output from this phase is number of cluster in addition to a number of connections not fit in any one of produced clusters this is known as outliers.

Attack summarization using association pattern: Discovered attacks and suspicious connections are summarized using Apriori algorithm presented by Jiawei and Kamber (2001). The algorithm describes the features detected outliers to assist the analyst in generating new signatures in the future.

EXPERIMENTS AND IMPLEMENTATION

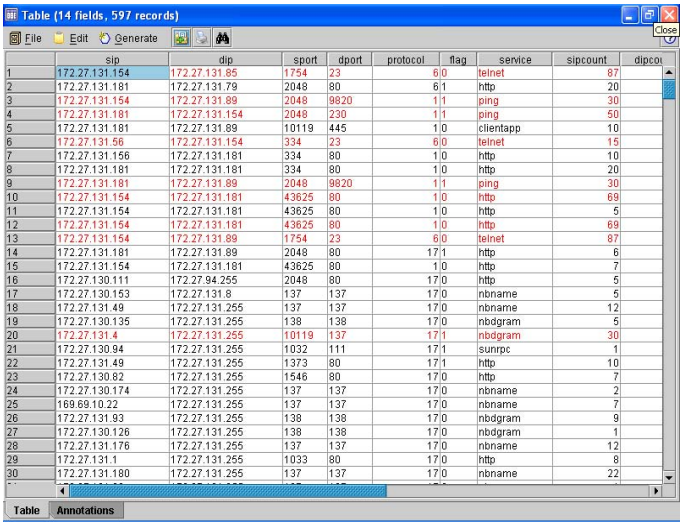
Using DARPA dataset to test the proposed framework: The proposed intrusion detection system is firstly applied to a small sample of 1998 DARPA (Intrusion detection evaluation data). The training data represents 80% of labeled data for 1 day that includes a total of 200 records. The data contains different types of attacks such as: Guess, port-scan. Phf, rlogin, Rsh and rcp. Testing data represents 20% of data from the same day.

Results: The number of records produced from the classification process (misuse detection) which are classified as intrusions are 21 in contrast to 21 classified as normal from testing data (20% from wall data). The number of misclassified records is about 3 records. The classification accuracy is about 97%. At anomaly detection phase the set of normal (21 record) are used as input. Data is placed into single cluster (clustter1) containing 13 records in addition to outliers ($null$) containing 8 records can be further analyzed by security administrator. The accuracy of anomaly detection is about 95.8%.

Experiments on real network data: Experiments are performed for a medium dataset captured from different parts of the western campus of SUST network (College of Computer Science and Information Technology-CCSIT-Network). Capturing has been performed during 90 min in to provide a semi complete image for types of data exchanged through the network. It contains 30 records. After filtering and feature extraction phases the total number of records is 3000 records. Collected data are further divided into training and testing with a certain percentage of 80% for training and 20% for testing. The reached results emphasis that decision tree algorithm is capable of discovering known attacks with accuracy reached to 99.6% for misuse detection phase. The Total number of records fed as a testing data to known attack detection model is 597 record. The output is classified into 2 classes. Normal class contains 565 records and intrusive class contains 32 records. The actual number of intrusive records is about 31 records and the actual number of normal records is 564 records. So there are 2 misclassified records 1 of them is belonging to actual normal set but it classified as intrusion. This is known as False Positive (FP) and the other 1 is belonging to actual intrusive set but it classified accidentally to normal. This is called False Negative (FN). So the number of false alarms produced from misuse detection model is only 1 record. Table 1 is a standard metrics for evaluations of intrusions (attacks) which shows the actual number of records in each

Table 1: Standard evaluation metrics for intrusion

	Predicted connection label	
Confusion matrix (Standard metrics)	Normal	Intrusions (attacks)
Actual connection label		
Normal	564 (TN)	1 (FP)
Intrusion (attacks)	1 (FN)	31 (TP)



	sip	dip	sport	dport	protocol	flag	service	sipcount	dipcou
1	172.27.131.154	172.27.131.85	1754	23	6	0	telnet	87	
2	172.27.131.181	172.27.131.79	2048	80	6	1	http	20	
3	172.27.131.154	172.27.131.89	2048	9820	1	1	ping	30	
4	172.27.131.181	172.27.131.154	2048	230	1	1	ping	50	
5	172.27.131.181	172.27.131.89	10119	445	1	0	clientapp	10	
6	172.27.131.56	172.27.131.154	334	23	6	0	telnet	15	
7	172.27.131.156	172.27.131.181	334	80	1	0	http	10	
8	172.27.131.181	172.27.131.181	334	80	1	0	http	20	
9	172.27.131.181	172.27.131.89	2048	9820	1	1	ping	30	
10	172.27.131.154	172.27.131.181	43625	80	1	0	http	69	
11	172.27.131.154	172.27.131.181	43625	80	1	0	http	5	
12	172.27.131.154	172.27.131.181	43625	80	1	0	http	69	
13	172.27.131.154	172.27.131.89	1754	23	6	0	telnet	87	
14	172.27.131.181	172.27.131.89	2048	80	17	1	http	6	
15	172.27.131.154	172.27.131.181	43625	80	1	0	http	7	
16	172.27.130.111	172.27.94.255	2048	80	17	0	http	5	
17	172.27.130.153	172.27.131.8	137	137	17	0	nbname	5	
18	172.27.131.49	172.27.131.255	137	137	17	0	nbname	12	
19	172.27.130.135	172.27.131.255	138	138	17	0	nbdgram	5	
20	172.27.131.4	172.27.131.255	10119	137	17	1	nbdgram	30	
21	172.27.130.94	172.27.131.255	1032	111	17	1	sunrpc	1	
22	172.27.131.49	172.27.131.255	1373	80	17	1	http	10	
23	172.27.130.82	172.27.131.255	1546	80	17	0	http	7	
24	172.27.130.174	172.27.131.255	137	137	17	0	nbname	2	
25	169.69.10.22	172.27.131.255	137	137	17	0	nbname	7	
26	172.27.131.93	172.27.131.255	138	138	17	0	nbdgram	9	
27	172.27.130.126	172.27.131.255	138	138	17	0	nbdgram	1	
28	172.27.131.176	172.27.131.255	137	137	17	0	nbname	12	
29	172.27.131.1	172.27.131.255	1033	80	17	0	http	8	
30	172.27.131.180	172.27.131.255	137	137	17	0	nbname	22	

Fig. 3: Discovered attacks highlighted with red color

class cross the predicted one. Figure 3 shows the output set of records from known attacks detection model. Detected attacks are highlighted.

The set of records classified as normal from the previous step are taken as input to anomaly detection phase. The number of input records is 565. The resulting out put contains two clusters in addition to outliers. Outliers represent a set of not classified records according to distance measures calculated by two-step clustering algorithm. The resulting set of outliers are viewed as suspicion connections and fed to association pattern analysis. Summery description of outliers are produced which assists security administrator to generate new signatures or even label.

EVALUATION OF THE PROPOSED FRAMEWORK

The proposed structure has been evaluated in order to measure its applicability and strength. Data mining accuracy measures related to the used algorithms is conducted. Also, standard evaluation metrics used to estimate the actual number of wrong classified records for known attack detection module. Thus, the overall evaluation confirms and ensures the feasibility and robustness of the proposed framework.

The proposed framework vs. MINDS: As mentioned previously MINDS has two major phases, misuse detection using signature based tool (i.e., SNORT) and anomaly detection using data mining technique. Since the proposed approach employs data mining for both detections, the comparison with the MINDS approach can be accomplished as follows: SNORT vs. decision tree C5.0 and two-step clustering algorithm vs. outlier LOF algorithm for the anomaly detection.

In misuse detection, the decision tree algorithm (C5.0) achieves high detection accuracy for known attacks reached to 99.6% vs. SNORT used in MINDS. Moreover, data mining is adaptive in nature, classification algorithm can generates rules automatically after passing network traffic to enable administrator to make sure with the correctness of the rules comparing to signature based tool that requires manually updating rules by human analysts each time a new suspicious behaviour is detected.

In anomaly detection, both proposed approaches use data mining. Ertoz *et al*. (2004) reported that in MINDS they are not able to testify the detection rate and false alarm rate due to difficulty in obtaining the complete labeling of network connections. However, according to some experiments in real network traffic at university of Minnesota, data mining based anomaly detection achieved success in discovering many novel network attacks and emerging network behavior that could not be detected using signature based systems such as SNORT. Conclusively such evaluation ensures that data mining can be used effectively with both misuse and anomaly detections and with high level of accuracy.

CONCLUSION

The obtained results show that using data mining techniques such as decision tree (C5.0 algorithm) and distance based clustering algorithms (Two-step clustering algorithm) are capable of discovering known and new attacks with acceptable level of false alarms. So it can be used effectively by security administrators to discover new emergent attacks from time to time. One of the potential drawbacks of using data mining is that it works only in an offline environment. The area of using data mining approaches for intrusion detection is an on-going

research area. Studying the pros and cons of the various techniques and choosing instance to implement or use is challenging issue in such context due to the variety of associated factors and parameters.

RECOMMENDATIONS

A number of issues could be done in future research:

- The nature of attacks that dynamically changed require real time detection using agent based intrusion detection system
- Using bigger datasets for testing to obtain more accurate results
- Refining the technique of coming up with a good threshold to improve detection accuracy
- Improve NIDS to support attacks prevention rather than only detection
- There is a need for visualization tool for providing a graphical user interface that helps security analysts to better comprehend the anomalous events and patterns extracted

REFERENCES

Brugger, T., 2004. Data mining methods for network intrusion detection. Technical Report. University of California, Davis. http://bruggerink.com/~zow/GradSchool/brugger_dmnid_survey.pdf.

Dokas, P. and L. Ertoz, 2002. Data mining for network intrusion detection. Proceedings of the NFS workshop on next generation data mining. Nov. 1-3, Marriott, Inner Harbor, Baltimore, pp: 21-29.

Dokas, P., L. Ertoz, V. Kumar, A. Lazarevic, J. Srivastava and P. Tan, 2002. Data Mining for Network Intrusion Detection. Proc. NSF Workshop on Next Generation Data Mining, Baltimore, MD.

Dunham, M.H., 2003. Data Mining Introductory and Advanced Topics. Prentice Hall of India, New Delhi, India, ISBN-10: 81-7758-880-X.

Eid, M., 2004. A new mobile agent-based intrusion detection system using distributed sensors. In proceeding of FEASC.

Ertoz, L., E. Eilertson, A. Lazarevic, A. Lazarevic and P. Tan, 2004. MINDS-minnesota intrusion detection system. Technical Report at University of Minnesota, pp: 1-21. http://static.msi.umn.edu/rreports/2005/68.pdf.

Giacinto, G. and F. Roli, 2002. Intrusion detection in computer networks by multiple classifier systems. Proceedings of ICPR 2002, 16th International Conference on Pattern Recognition, Quebec City, Canada. Aug 11-15, IEEE press, pp: 390-393.

Idris, N.B. and B. Shanmugam, 2006. Novel attack detection using fuzzy logic and data mining. Security and Management, pp: 26-31. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.8373&rep=rep1&type=pdf.

Jiawei, H. and M. Kamber, 2001. Data Mining: Concepts and Techniques. Higher Educa-tion Press, Beijing, China, pp: 3-10.

Lee, W. and S. Stolfo, 1998. Data mining approaches for intrusion detection. Proceeding of the 7th USENIX sec. Symposium, San Antonio, Texas, Jan. 26-29, USENIX Association, Berkeley, CA, USA., pp: 1-16.

Lee, W., S.J. Stolfo and K.W. Mok, 1998. Mining audit data to build intrusion detection models. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, Aug. 27-31, AAAI Press, New York, pp: 1-20.

Lee, W., S.J. Stolfo and K.W. Mok, 1999. A data mining framework for building intrusion detection models. http://www.cs.earlham.edu/~aburdma/154973.html.

Peng, T. and W. Zuo, 2006. Data mining for network intrusion detection system in real time. Int. J. Comput. Sci. Network Sec., 6: 173-177.

Prasad, G.V.S.N.R.V., Y. Dhanalakshmi, V.V. Kumar I.R. Babu, 2008. Modeling an intrusion detection system using data mining and genetic algorithms based on fuzzy logic. IJCSNS Int. J. Comput. Sci. Network Sec., 8: 319-325.

Shanmugam, B. and N.B. Idris, 2009. Improved intrusion detection system using fuzzy logic for detecting anamoly and misuse type of attacks. Proceeding of International Conference of Soft Computing and Pattern Recognition, Dec. 4-7, IEEE, pp: 212-217.

Shyu, M.L. and V. Sainani, 2009. A Multiagent-Based Intrusion Detection System with the Support of Multi-Class Supervised Classification. In: Data Mining and Multi-Agent Integration, Cao, L. (Ed.). Springer-Verlag, USA., Australia, pp: 127-142.

Zhu, D., G. Premkumar, X. Zhang and C.H. Chu, 2001. Data mining for network intrusion detection: A comparison of alternative methods. Decision Sci., 32: 635-660.