

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & MANAGEMENT

INTRUSION DETECTION SYSTEM BASED ON DB SCAN AND SUPPORT VECTOR MACHINE

Prateek Rokadiya^{*1}, Saumy Dadhich², Sandesh Joshi³, Akshita Pamecha⁴, Payal Dodeja⁵, Anurag Punde⁶

ABSTRACT

In this paper, we will discuss about the intrusion detection system their effects and also discuss about the existing algorithms like clustering algorithms such as K means their drawbacks and effects. We will also discuss one more algorithm called DBSCAN (Density based scanning algorithm). We have compared this algorithm with the existing algorithm and present the summary also.

Keywords- DBSCAN, clustering, K-means, f-measure, efficiency.

I. INTRODUCTION

We first employ a clustering algorithm to partition a raining data set that consists of labeled flows combined with unlabeled flows. Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. A K-Means clustering algorithm attempts to find natural groups of components (or data) based on some similarities. The K-Means clustering algorithm also finds the centroid of a group of data sets. The k-means algorithm used in this work is one of the most non-hierarchical methods used for data clustering.

After clustering of training data, we use the available labeled flows to obtain a mapping from the clusters to the different known classes. The result of the learning is a set of clusters, some mapped to the different flow types. This method, referred to as semi-supervised learning. The input data for classification task is collection of number of records. Each record, also known as an instance, is characterized by a tuple (x, y), where x is the attribute set and y is class attribute. Let $X = \{X_1 \dots X_N\}$ be a set of flows. A flow instance X_i is characterized by a vector of attribute values, $X_i = \{X_{ij} | 1 \leq j \leq m\}$, where m is the number of attributes, and X_{ij} is the value of the jth attribute of the ith flow. Also, let $Y = \{Y_1 \dots Y_q\}$ be the set of traffic classes, where q is the number of classes of interest. The Y_i 's can be classes. Our goal is to learn a mapping from an m-dimensional variable X to Y. This mapping forms the basis for classification models. This way the trained system is formed and it is then tested. In testing stage, after the training phase is over next is to test it on out-of-sample data. The testing phase is basically depends on the result parameters of training phase. In testing phase minimum distance of each record from cluster center obtained from the training phase is compared, If found the data is assigned the same cluster. But, there are certain problems related with K-means clustering. These are as follows:

- Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome).
- Fixed number of clusters can make it difficult to predict what K should be.
- Does not work well with non-globular clusters.
- Different initial partitions can result in different final clusters. It is helpful to rerun the program using the same as well as different K values, to compare the results achieved.

II. PROBLEM STATEMENT

There have many classifiers to find the intrusion, such as Clustering classifier, neural network classifier, Bayesian classifier and SVM classifier. It has been observed that to select correct classifier is a tough work.

Although many IDS has been developed many years ago, but it generate large amount of alert messages which makes the maintenance of system inefficient. Most of the presented IDs make use of all the features in the packet to analyze and look for well - known intrusive model. Some of these features are irrelevant and redundant.

Existing approaches have the following drawbacks.

- Lengthy training and testing process,
- Low accuracy rate,
- High false positive rate,
- Low detection rate and
- Occupy more storage space.

Apart from these K-Means clustering have the many drawbacks like the clusters are of differing sizes, densities and non globular shapes. It also has the problems with outliers and it generates empty clusters.

Proposed approach finds out the problems and removed the drawbacks of existing approaches. DB SCAN is one of the promising algorithms for intrusion detection system.

In the proposed approach, we have used Support Vector Machine as a classifier.

III. PROPOSED APPROACH

In previous chapter, the most regularly used intrusion detection techniques are described. Existing techniques are used the Clustering classifier, neural network classifier and Bayesian classifier. These approaches have the problem of over-fitting. SVM classifier removes the problem of over-fitting, but it uses the all features or maximum features of dataset to find the accuracy rate, so there are increases the problem of data redundancy and it consumes more computer recourses. To remove the drawbacks of existing approaches, DBSSVM approach is proposed. In this proposed approach, Intrusion Detection System is implemented by using Rough Set Theory and Support Vector Machine technique. RST is used for feature selection, and SVM is used as classifier in this system.

In this work, accuracy rate, false positive rate and attack detection rate of intrusion detection using rough set theory and support vector machine are tried to find out. Since SVM classifier supports only numeric data, so in the proposed approach, firstly we have to convert the text features in numeric data, then we get the small size dataset by reducing the redundant data. Finally, selected features are passed to the SVM to get the accuracy rate, false positive rate and attack detection rate of dataset. The proposed algorithm has following steps.

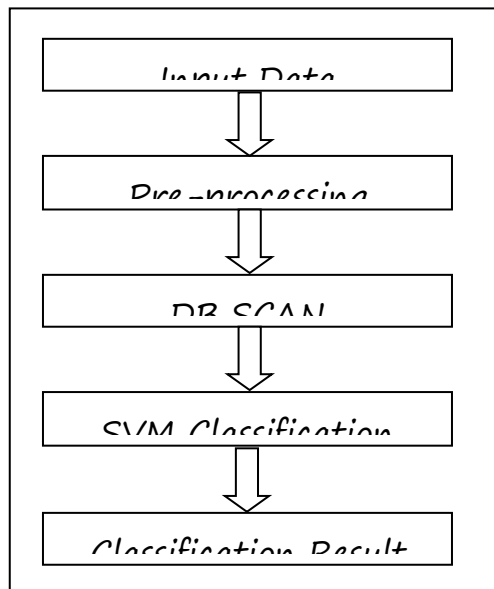
- Data Preprocessing.
- DB SCAN Algorithm.
- Intrusion SVM Classification

IV. ARCHITECTURE OF PROPOSED APPROACH

Intrusion detection is a method for finding various intrusions. But here we are trying to find their behavior. As there are various factors on which we can evaluate the exact behavior of an intruder. But in earlier researches, some factors give same type of information or we can say that redundant information is obtained. So this becomes a very tedious job to rectify such redundant information. It took more space in computer memory.

Intrusion detection is a critical component of secure information systems. Since elimination of the insignificant and/or useless inputs leads to a simplification of the problem, faster and more accurate detection may result. DB SCAN, therefore, is an important issue in intrusion detection.

So, to overcome from such type of redundancy and to identify the important features, an approach is proposed and implemented in this dissertation. By using DB SCAN algorithm, we are able to minimize the 41 features of KDD CUP'99 dataset into 6 features. These selected 6 features are non-redundant and give focused information. So this is a step towards more accuracy. Then support vector machine tool is used to classify the data, and find the accuracy of detection. Proposed method follows the necessary steps required to perform in Intrusion Detection System. These steps are Data Preparation, Data Preprocessing, Feature Selection, and Classification. The flow chart of proposed DBSSVM approach for Intrusion Detection is illustrated in Figure 4.1



Proposed DBSSVM Approach for ID

Since standard (benchmark) dataset for intrusion detection is available so there is no need to prepare the dataset. KDD CUP'99 dataset is used as a database to test the system performance, which is the dataset used for the third

international knowledge discovery and data mining tools competition, which was held in conjunction with KDD CUP'99, the fifth international conference of knowledge discovery and data mining. For selecting features from each sample, rough set theory has been adopted. These selected features corresponding to each instance proceed for classification. As a classification technique, SVM is chosen because it already provided a better accuracy than other techniques for intrusion detection.

V. DATASET DESCRIPTION

Since 1999, KDD'99 has been the most widely used data set for the evaluation of anomaly detection methods. The database is gathered from The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99, The Fifth International Conference on Knowledge Discovery and Data Mining. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. The simulated attacks fall in one of the following four categories.

- **Denial of Service Attack (DoS):** DoS is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.
- **User to Root Attack (U2R):** U2R is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.
- **Remote to Local Attack (R2L):** R2L occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.
- **Probing Attack:** Probe is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

Table 4.1: Attack types and their respective classes

Attack Type	Attack Category	Attack Type	Attack Category	
ftp write	R2L	apache2	DoS	
guess_passwd		Back		
Imap		Land		
Multihop		Mailbomb		
Named		Neptune		
Phf		Pod		
Sendmail		Processtable		
Snmpgetattack		Smurf		
Snmpguess		Teardrop		
Spy		Udpstrom		
Warezclient		Normal		Normal
Warezmaster		buffer_overflow		U2R
Worm		Httpunnel		
Xlock		Loadmodule		
Xsnoop	Perl			
Ipsweep	Ps			
Mscan	Rootkit			
Nmap	Sqlattack	Probe		
PortswEEP	Xtern			
Satan				
Saint				

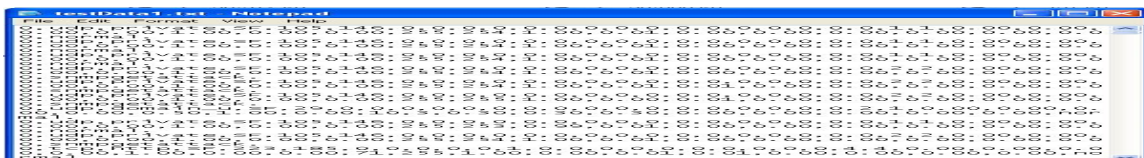


Figure 4.2: The original data of KDD CUP'99

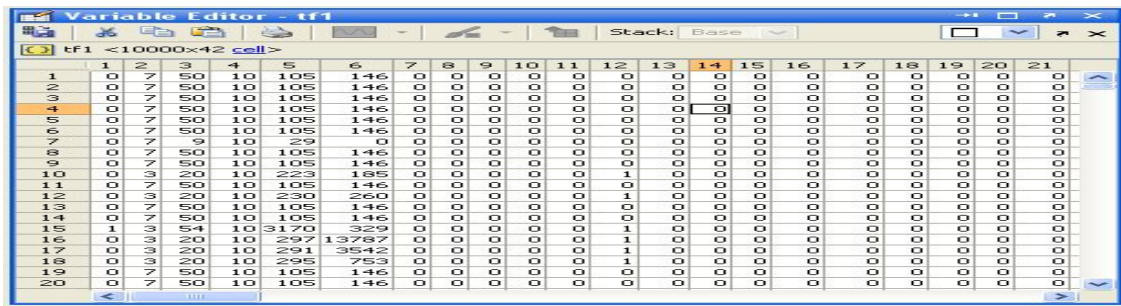
A. DATA PREPROCESSING

The information obtained by KDD Cup’99 is a combination of many system calls. A system call is a text base record. Every system call in the dataset has 41 features as listed in table 4.2. There are several text words in the dataset. Since SVM is used only numerical data for testing and training, so text features are needed to be converted into numerical values. The features, as shown in figure 4.1, contain the text value are protocol_type, flag, and service. Therefore, I have assumed some numerical values for different text features, like ‘protocol_type’ feature ‘tcp’ as 3, ‘udp’ as 7, and ‘icmp’ as 9 etc. To translating the Text data to numeric data in KDD cup’99 Data Set is given in table 4.2.

Table 4.3: Transformation Table to convert Text to Numeric

Type	Class	No.	Type	Class	No.
Attack/ Normal	Attack	1	Service	imap4	23
	Normal	0		iso_tsap	24
Protocol Type	TCP	3		Klogin	25
	UDP	7		Kshell	26
	ICMP	9		Ldap	27
Flag	OTH	1		Link	28
	REJ	2		Login	29
	RSTO	3		Mtp	30
	RSTOS0	4		Name	31
	RSTR	5		netbios_dgm	32
	S0	6		netbios_ns	33
	S1	7	netbios_ssn	34	
	S2	8	Netstat	35	
	S3	9	Nnsp	36	
	SF	10	nntp	37	
	SH	11	telnet	38	
Service	Auth	1	Time	39	
	Bgp	2	Uucp	40	
	Courier	3	uucp_path	41	
	csnet_ns	4	Vmnet	42	
	Ctf	5	Whois	43	
	Daytime	6	Z39_50	44	
	Discard	7	ntp_u	45	
	Domain	8	Other	46	
	domain_u	9	pop_2	47	
	Echo	10	pop_3	48	
	eco_i	11	Printer	49	
	ecr_i	12	Private	50	
	Efs	13	remote_job	51	
	Exec	14	Rje	52	
	Finger	15	Shell	53	
	ftp	16	Smtplib	54	
	ftp_data	17	sql_net	55	
	Gopher	18	Ssh	56	
	Hostnames	19	Sunrpc	57	
	http	20	Supdup	58	
	http_443	21	Systat	59	
	IRC	22	X11	60	

After transformation of text values to numeric value, the dataset is obtained in following format. Features of KDD CUP'99 and behavior of intruder is shown in figure 4.3.



Continue.....

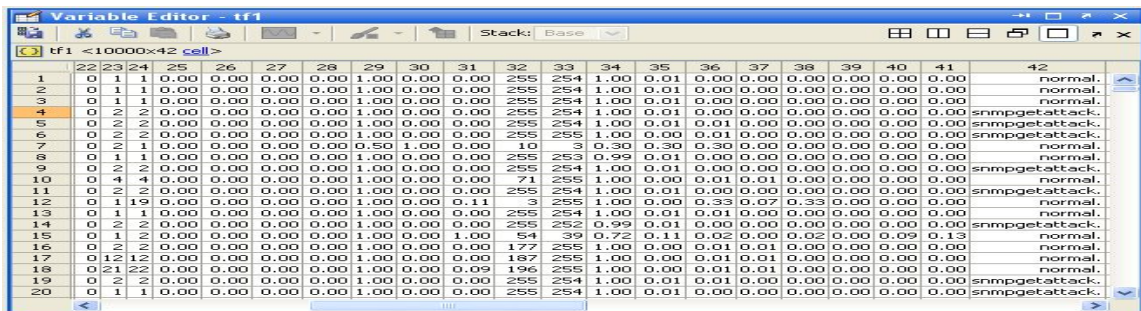


Figure 4.3: 41 Features and behavior of Intruder

VI. RESULT ANALYSIS

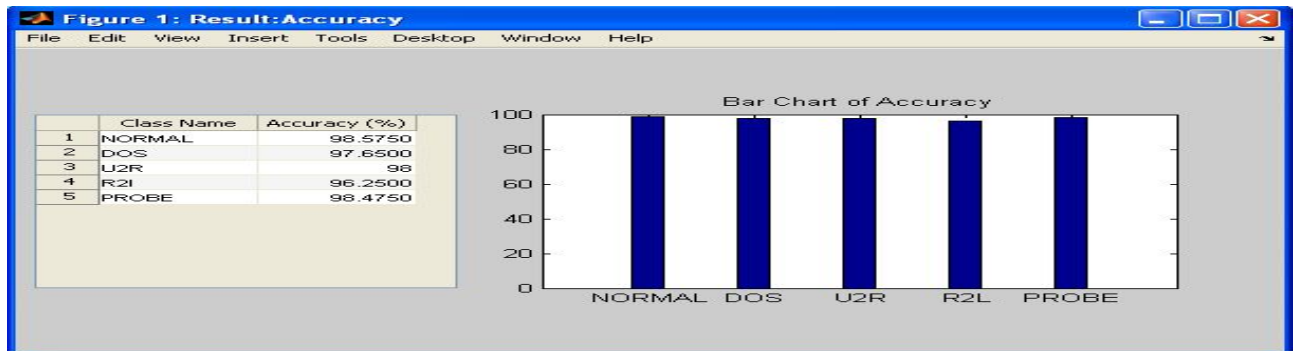


Figure 5.1 Analysis of Accuracy Rate

PRECISION

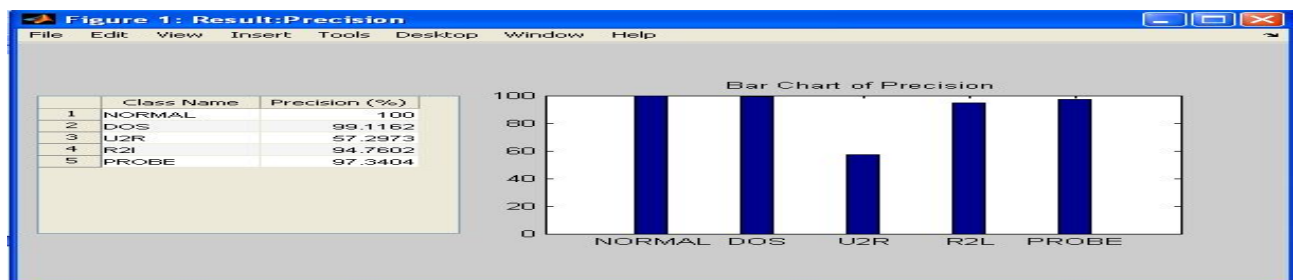


Figure 5.2: Analysis of precision

RECALL

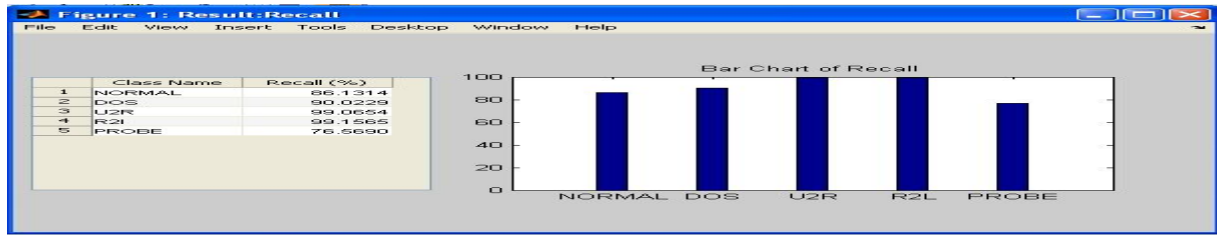


Figure 5.3: Analysis of Recall

F-MEASURE

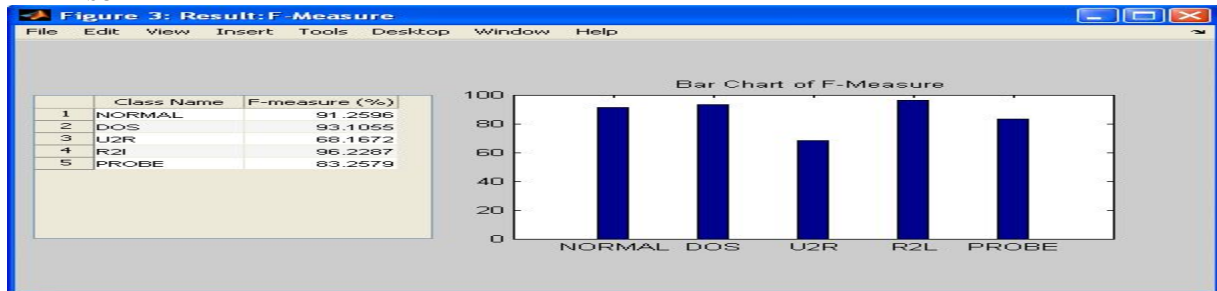


Figure 5.4: Analysis of F-Measure

CONFUSION MATRIX

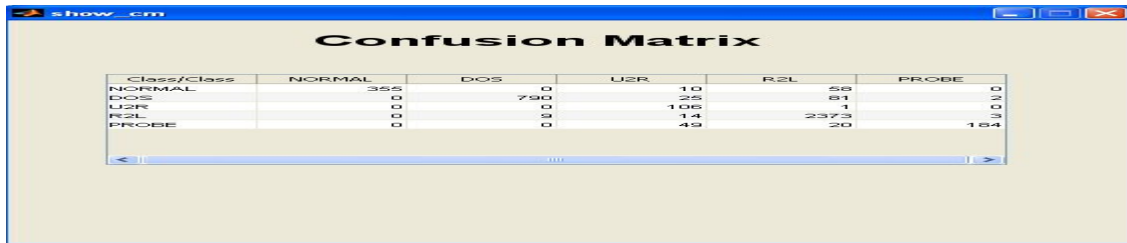


Figure 5.5: Confusion Matrix

CLUSTERING RESULT

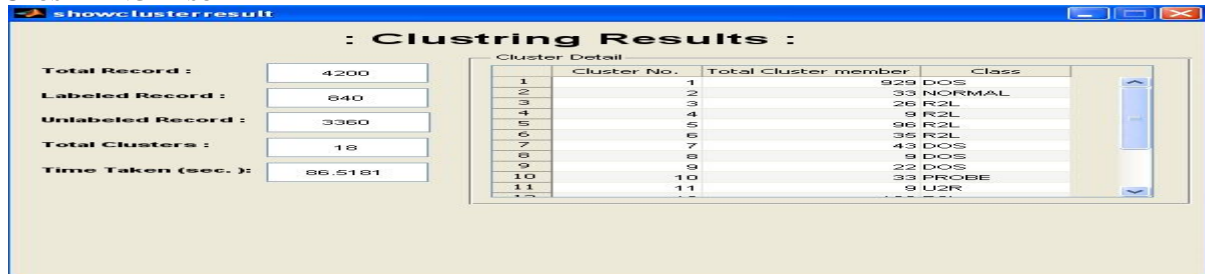


Figure 5.6: Clustering Result

RESULT FOR KDD DATA SET

Attack Type	Accuracy Rate	Precision	Recall	F measure
Normal	98.5750	100	86.1314	91.2596
Dos	97.6500	99.1162	90.0229	93.1055
U2R	98	57.2973	99.0654	68.1672
R2L	96.2500	94.7602	99.1565	96.2787
Probe	98.4750	97.3404	76.5690	83.2579

TABLE 4.4 CLUSTERING RESULT FOR KDD CUP DATA SET

Training Data Set	No. of Features	Labeled Samples	Unlabeled Samples	No. of Clusters	Time Taken (Sec)
4200	41	840	3360	18	86.5181

VII. CONCLUSION

The aim of the project is to design and implement a semi-supervised learning approach for network traffic classification and it has been achieved successfully. A DB SCAN approach to design a Network Traffic Classifier is implemented successfully. Algorithm permits both labeled and unlabeled data to be used in training the network. While performing training and testing of the classifier for a dataset, it is observed that a test error rate depends on the number of clusters which is randomly used in training phase. We have used the KDD Data Set as a training data set and improved the accuracy rate of the semi supervised algorithm.

Proposed algorithm is very apt and reliable for finding the supervised and unsupervised data. The algorithm has been proved that in the future Of course, we can improve accuracy rate, false positive rate and attack detection rate of intrusion detection by providing the some improved form of the DB SCAN algorithm.

REFERENCES

1. J.A. Hartigan (1975). *Clustering algorithms*. John Wiley & Sons, Inc.
2. Hartigan, J. A.; Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 28 (1): 100–108. JSTOR 2346830
3. S. V. Sabnani, "Computer Security: A Machine Learning Approach", 2008.
4. A. Patcha, J-M. Park, "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends", *Computer Networks* (2007).
5. Denning, Dorothy E., "An Intrusion Detection Model," *Proceedings of the Seventh IEEE Symposium on Security and Privacy*, May 1986, pages 119–131.
6. NIST – *Guide to Intrusion Detection and Prevention Systems (IDPS)*". 2007-02. Retrieved 2010-06-25
7. W. Li, "Using Genetic Algorithm for Network Intrusion Detection," C. S. G. Department of Energy, Ed., 2004, pp. 1-8.
8. J. Macqueen, *Some methods for classification and analysis of multivariate observations*," *Proc. of the Fifth Berkeley Symp. On Math. Stat. and Prob.*, vol. 1, pp. 281-296, 1967.
9. E. Forgy, *Cluster analysis of multivariate data: efficiency vs interpretability of classification*," *Biometrics*, vol. 21, pp. 768, 1965.
10. D. J. Hall and G. B. Ball, *ISODATA: A novel method of data analysis and pattern classification*," *Technical Report, Stanford Research Institute, Menlo Park, CA*, 1965.
11. The history of k-means type of algorithms (LBG Algorithm, 1980) R.M. Gray and D.L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, Vol. 44, pp. 2325-2384, October 1998. (Commemorative Issue, 1948-1998)
12. Martin Ester, Han-peter Kriegel, Jorg Sander, Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", 2nd International conference on Knowledge Discovery and Data Mining (KDD-96).
13. Judy Weng, "Network Intrusion Prevention Systems", *JTB Journal of Technology and Business*. October 2007.
14. IBM System, *i5/OS Information Center I Security Intrusion detection Version 5 Release 4*, 2007.
15. A. H. Sung, S. Mukkamala, "Identifying Important Features for Intrusion Detection Using Support Vector Machines and Neural Networks," in *2003 Symposium on Applications and the Internet 2003*, pp. 209-216.
16. C. Endorf, et al, "Intrusion Detection & Prevention", ISBN: 0072229543, McGraw-Hill, 2004.
17. R. P. Lippmann, et al, "Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation," *disceX*, vol. 02, p. 1012, 2000.
18. Andreas Vlachos, "Active Learning with Support Vector Machines", *School of Informatics University of Edinburgh* 2004.
19. *Network Security by Williams Stalling*.
20. R. C. Holte, "Very simple classification rules perform well on most commonly used datasets." *Machine Learning*, 11(1):63–91, Apr. 1993.
21. Wang, Ke. "Anomalous Payload-Based Network Intrusion Detection". *Recent Advances in Intrusion Detection*. Springer Berlin. doi:10.1007/978-3-540-30143-1_11. Retrieved 2011-04-22.
22. A strict anomaly detection model for IDS, *Phrack* 56 0x11, Sasha/Beetle
23. Perdisci, Roberto; Davide Ariu, Prahlad Fogla, Giorgio Giacinto, and Wenke Lee (2009). "McPAD : A Multiple Classifier System for Accurate Payload-based Anomaly Detection". *Computer Networks, Special Issue on Traffic Classification and Its Applications to Modern Networks* 5 (6): 864–881.
24. S. Forrest, S. A. Hofmeyr, A. Somayaji, and T. A. Longstaff, "A Sense of Self for Unix Processes," in *IEEE Symposium on Research in Security and Privacy*, Oakland, CA, USA, 1996, pp. 120--128.

24. C. Warrender, et al, "Detecting Intrusions Using System Calls: Alternative Data Models," in *IEEE Symposium on Security and Privacy*, Oakland, CA, USA, 1999, pp. 133-145.
25. D. Heckerman, "A Tutorial on Learning with Bayesian Networks," *Microsoft Research, Technical Report MSR-TR-95-06*, March 1995.
26. A. Valdes, K. Skinner, "Adaptive Model-based Monitoring for Cyber Attack Detection," in *Recent Advances in Intrusion Detection Toulouse, France*, 2000, pp. 80--92.
27. R. A. Calvo, et al, "A Comparative Study of Principal Component Analysis Techniques," in *Ninth Australian Conference on Neural Networks*, Brisbane, Queensland, Australia, 1998.
28. M.-L. Shyu, et al, "A Novel Anomaly Detection Scheme Based on Principal Component Classifier," in *IEEE Foundations and New Directions of Data Mining Workshop*, Melbourne, Florida, USA, 2003, pp. 172-179.
29. N. Ye, Y. Z. C. M. Borrer, "Robustness of the Markov-Chain Model for Cyber-Attack Detection," *IEEE Transactions on Reliability*, vol. 53, pp. 116-123, March 2004.
30. Arun Pujari, "Data Mining Concepts", pp 1-5.
31. J. E. Dickerson, J. A. Dickerson, "Fuzzy network profiling for intrusion detection," in *19th International Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, Atlanta, GA 2000, pp. 301 – 306.