Protocol Based Network Intrusion Detection System

Umanand Pandey

Master of Computer Applications

J.S.S. Academy of Technical Education,Noida(UP)

---***---

Abstract – Considering the todays scenario of development, the new and more flexible network based applications are being developed, so we need a lightweight and fast accessing Network Intrusion Detection System. By eliminating the inefficient/useless or common features it can obtain a higher performance in the terms of Intrusion Detection, to do that here Logistic Regression has been used. In this paper study, KDDCUP'99 has been used as the dataset for evaluation purpose.

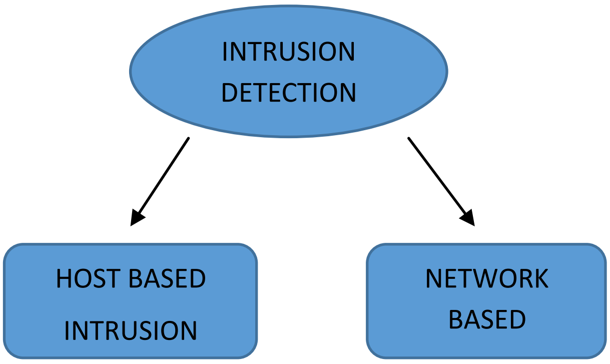
Key Words: Intrusion Detection, Protocol, Logistic Regression (LR), Support Vector Machine (SVM)

1.INTRODUCTION

Intrusion detection system can be a hardware device or any software application that can sense a malicious activity over a network. The malicious activity involves the unauthorized access of data, or violation of recommended policies over the network. Now a days the world is running after digitalization of every single activity that can be performed over network. In the rush of digitalization of critical activities over the network, it is necessary to ensure the security of the information from threatening breaches. The Intrusion Detection System looks for dubious network traffic and generates interrupts when such activity is encountered. IDS (Intrusion Detection System) scans a network or system to figure out malicious activity or policy violations. Malicious activity detected by IDS is reported either to the admin or saved centrally using a Security Information and Event Management system (SIEM). The SIEM system combines input from multiple sources and uses alarm sorting techniques to distinguish hazardous activities from false alarms.

Intrusion detection is the process of monitoring events occurring in a computer system or network and analyzing them for signs of intrusions [1].

IDS are of two types:



Host Based Intrusion Detection System (HIDS) runs over data packets received from a particular host and then store raw networks packets in the form of source of data from the network and recognize the footprints of intruders. A HIDS looks into the traffic to the host and from the host over the system IDS is running. A HIDS has the ability to monitor key system files and any attempt to overwrite these files.

NIDS helps an organization to look after the cloud services, on host system and cloud or remote services over the malicious activities which can come through the threatening of data. It contains policy violations, port scanning, and unidentified source and destination traffic. NIDS is a technology that is 'passive' rather than 'active' in nature. NIDSs are developed to alert on malicious events, and because of that are evenly run along with IPS (Intrusion Prevention Systems) that are 'active'. NIDS collects information about outgoing and incoming internet traffic. The sensors of NIDS are used strategically to increase visibility, across the network, e.g. on a LAN and DMZ. Basically NIDS combines the two methods 1) signature-based detection 2) anomaly-based detection. In the first method i.e. signature-based detection, the contents of collected data packets are compared against the signature files that are previously recognized as vulnerable. Whereas the second method (i.e. anomaly-based detection), uses behavioral analysis against a baseline of 'typical' network activity to monitor the events. NIDS limitations.

2. LITERATURE SURVEY

Selection of feature may reduce the computation and data complexity. Some more efficient and useful feature subsets can be get using feature selection. The previous works in the field of anomaly detection carried out in different researches using different machine learning algorithms are discussed below:

SVM is a powerful algorithm of machine learning which provides the tools towards generalized class with insufficient candidate ability. It was proposed by the Russian statistician and mathematician Vladimir Naumovich Vapnik. Ambwani [2] used SVM as classification tool. In Ambwani [2] experiment, comparing against the KDDCUP'99 test dataset, the accuracy rate appeared 92.46%. It is pointed out that KDDCUP'99 test dataset does not facilitate in forecasting results [3]. KDDCUP'99 full dataset was used as test based on the proposed Ambwani theory, the prediction accuracy rate was 99.9382% [4].

Logistic Regression, another algorithm of machine learning and variant of regression analysis, is a statistical model that uses a logistic function to model a binary dependent variable. It was proposed by Verhulst, a mathematician from Belgium in 19th century. Since the number of variables in two dependent variable error categories cannot be strained with, so the Logistic Regression is used to solve traditional linear regression. There is a critical rate of increase (threshold) S-function with the highest probability (Maximum Likelihood Estimation; MLE) predictors of the best parameter estimates, which can account for two types of explicit variables that make the forecast more accurate. The LR was used as an IDS feature selection and the test data was the full details of KDDCUP'99. The correct rate was 99.95% [5].

Discriminant analysis is a mathematical method used to divide observations into overlapping groups, based on scores on one or more statistical variables. The discriminant analysis (DA) is used to detect discriminant validity between two or more naturally occurring groups. The DA works by creating discriminatory activities (DFs) that predict which party each party belongs to. DFs are interpreted with the same coefficients as the structural matrix. DFs create a boundary between groups. Wong used DA as the feature selection method and the false alarm rate was 0.37% in 9 selected features [6].

PCA simplifies eigenvector-based multivariate analyzes. Generally, its functionality can be thought of as expressing the internal structure of data in a way that best describes the diversity of data. Depending on the field of application, K. Person proposed Principal Component Analysis (PCA) [3]. It is also known as the discrete Karhunen-Loève transform. PCA is based on converting a large number of variables into a small number of unrelated variables by finding a few orthogonal linear combinations of variables that start with the largest variable. The 14 features were chosen to predict the accuracy which was 99.8734 of the KDD CUP'99 full data (kddcup.data.gz) [4].

To remove or neglect the discrete eigenvalue, to reduce the number of features replacing the original feature set the Discriminant Analysis and Principal Component Analysis are clustering features as well as simplifying feature subsets. Using Logistic Regression, the calculating complexity and the data dimensional complexity can be reduced by determining protocol as conditional to prevent the discrete Eigen value being ignored and maximize the efficiency of Logistic Regression. Thus, in this research work, Logistic Regression was chosen as the main method, and same was then compared against the methods for PCA [4], DA [5], LR [5].

Protocol Anomaly Detection (PAD) works by analyzing the service traffic level, commands and behavior and then blocking and denying undesirable or inappropriate orders. The application rules have been published in RFCs and vendor documents [7]. Application protocols can be used to determine appropriate or expected behavior, however unavailable; new attacks can be effectively prevented. 90% of the attacks are protocol usage anomalies. The reason for that is most of the attacks exploit breaches in badly defined areas of protocols both in the protocol standard itself as well as its implementations. For example, CodeRed used buffer overflow to determine attacks [8]. Thus, using communication protocol makes intrusion detection models more efficient.

3.METHODOLOGY

Protocol Anomaly Detection (PAD) works by analyzing the service traffic level, commands and behavior and then blocking and denying undesirable or inappropriate orders. Application protocols have been published in RFCs and vendor documents [7]. Application protocols can be used to determine appropriate or expected behavior, however unavailable; new attacks can be effectively prevented. According to [8], the 90% of attacks over network are protocol usage inconsistencies. The reason for that is the majority attack to exploit violations in poorly defined areas, both protocols in the standard protocol itself and its implementations.

In this work, communication protocols such as TCP, ICMP and UDP are used to practice on different protocol based Intrusion Detection System.

Whole workflow can be represented through following algorithm

3.1 Step 1: Preprocessing of data: The data from dataset KDDCUP 99 may contain redundancy missing value issues. To overcome these issues preprocessing of data is required.

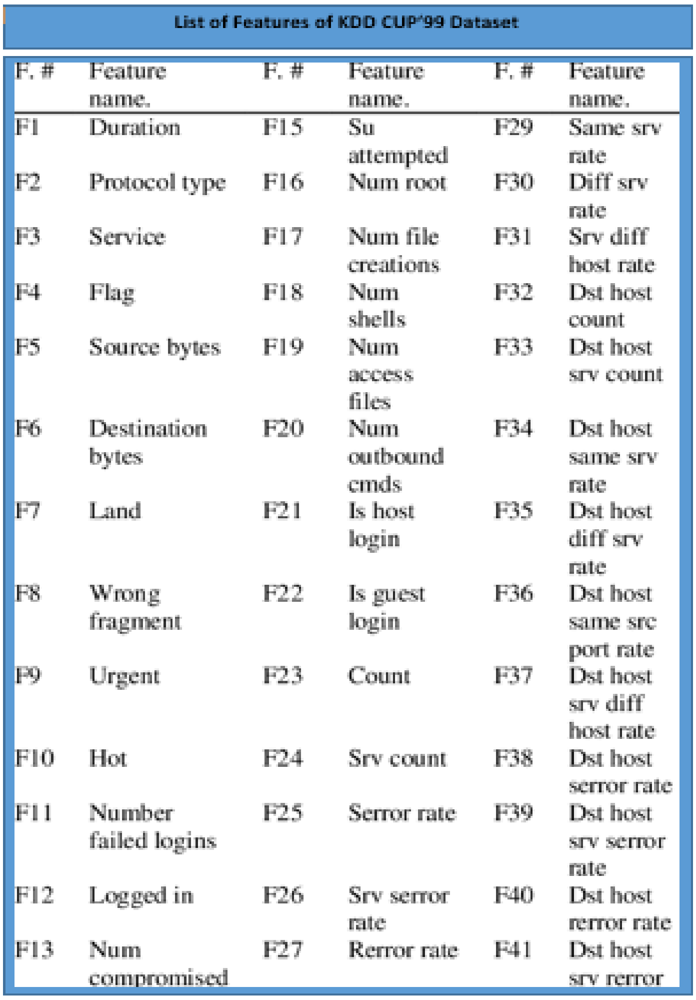
3.2 Step 2: Feature Selection and Distribution of Data: Feature selection process to reduce the amount of input variables while the establishment of a prediction model. It is desirable to reduce the amount of input switches to include both computer costs and, in some of the cases, to improve model performance. After preprocessing the data is ready to perform further steps.

Features of data are identified and divided into different groups on the basis of communication protocols. Data is needed to handle very carefully due to the minor difference in numerical values.

3.3 Step 3: Using SVM for Classification: The SVM can be used for the both classification and regression. The SVM uses a technique called the kernel trick to convert the data and based on this conversion only it finds the perfect boundary between possible outputs. In simple words, it makes data conversion more complex, and calculate how to use the data based on the labels or results that have defined. Thus SVM has been used to remove discrepancies in the classification of data. As this is well known that the rate of prediction of SVM classification is fine, there are the two important subjects that influences prediction, the one is the “kernel function selection” and the other one is the “hyper-parameters search”. Selecting the suitable kernel function and the best hyper-parameter is critical issue for SVM. But even the common way to solve this problem has been trial and error.

The four kernel functions are there: Radical Basic Function (RBF), Linear, Sigmoid and Polynomial. There is no any standard to choose a suitable Kernel function, but research of Smola says that, the general first reasonable choice is the RBF function [12]. The two necessary parameters in RBF Kernel are: C and γ that must be searched. So here the main goal is to identify better parameters (C, γ) so that the prediction of unknown data can be done using classifier accurately. Now, the scholar and the experts, in effort to get the solution for selection of the SVM parameter, some efforts have been proposed [10]. Chang, J. Lin, developed the LIBSVM which uses cross-validation parameters to achieve the best approach [11]. In addition to the above methods can be selected SVM best parameters, academics Ambwani proposes other solutions [2]. The only two parameters C and γ can be provided by the LIBSVM RBF kernel function, the first selected method as a numeric value of the static value of γ (γ LIBSVM default value of 1 / k, k values for the input attributes number [5]), and the parameters for C numerical interval t were set as an instable SVM forecast information and module training.

4. EXPERIMENT



List of Features of KDD CUP'99 Dataset

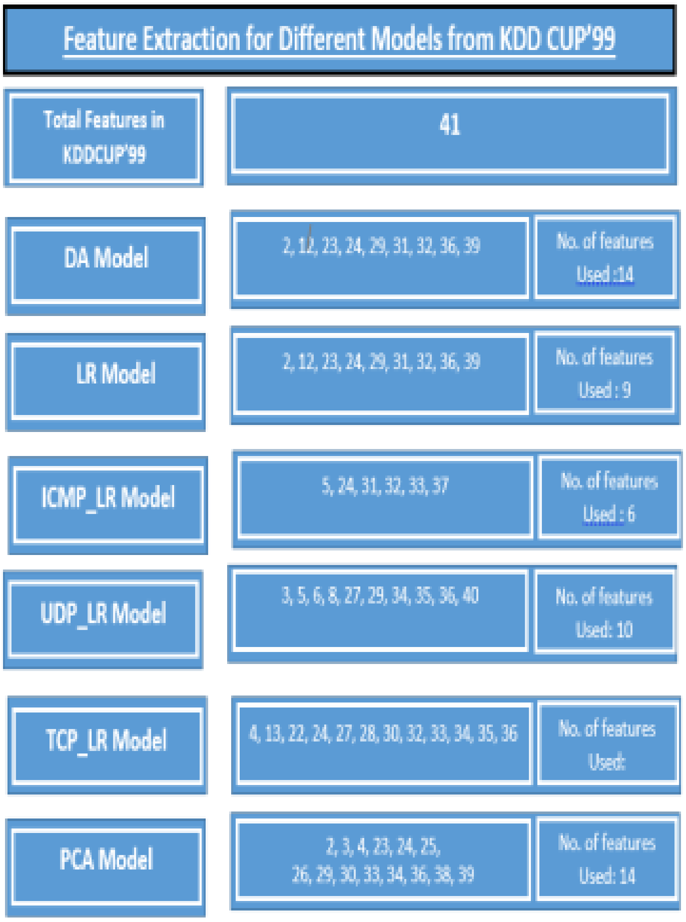
F. #	Feature name.	F. #	Feature name.	F. #	Feature name.
F1	Duration	F15	Su attempted	F29	Same srv rate
F2	Protocol type	F16	Num root	F30	Diff srv rate
F3	Service	F17	Num file creations	F31	Srv diff host rate
F4	Flag	F18	Num shells	F32	Dst host count
F5	Source bytes	F19	Num access files	F33	Dst host srv count
F6	Destination bytes	F20	Num outbound cmds	F34	Dst host same srv rate
F7	Land	F21	Is host login	F35	Dst host diff srv rate
F8	Wrong fragment	F22	Is guest login	F36	Dst host same src port rate
F9	Urgent	F23	Count	F37	Dst host srv diff host rate
F10	Hot	F24	Srv count	F38	Dst host serror rate
F11	Number failed logins	F25	Serror rate	F39	Dst host srv serror rate
F12	Logged in	F26	Srv serror rate	F40	Dst host rerror rate
F13	Num compromised	F27	Rerror rate	F41	Dst host srv rerror

Fig -1: KDDCUP ’99 Feature List

The experiments have been done over a Windows 10 machine with a Intel Core i3-6006U CPU @ 2.0GHz processor and 2GB ram. As the training dataset the kddcup.data_10_percent.gz with 494,020 records has been used and the 4,898,430 records, KDDCUP'99 as the complete dataset (kddcup.data.gz) used for the testing purpose. The concept of KDDCUP'99 from a research study done and post-processed by Columbia University at the MIT Lincoln Lab, Involving the four attack categories as: Dos Denial of service. In the early days, the KDDCUP'99 competition used the corrected.gz as test dataset, but according to the huge data difference will lead to poor detection accuracy [3].

To make the eigenvalue simple, the logistic regression is used stepwise, and for the data classification according to the diff. protocols (communication protocols), this study is used. The protocols are actually divided into the five different parts: feature selection using SPSS 13.0 statistical software, and validating the t and Support Vector Machine classification test. As we have studied above, to deduct the duplicity 20 features (to reduce the duplicate selection features in UDP_LR, TCP_LR and ICPM_LR) to setup the models with diff Communication Protocols. The figure-2 shows best selection feature subsets.

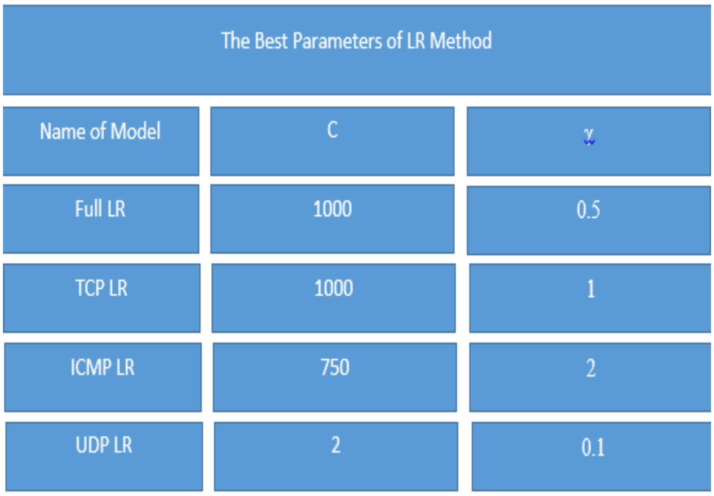
For the training purpose the feature sets of extracted dataset has been fed to the Support Vector Machine and the dataset for the testing purpose has been examined prior to the process of training been has done. It must determine, the parameters, C and γ, of the Gaussian Radial Basic Function (RBF). To find the best result yielding parameters, to train the dataset, the 10-fold CV (Cross Validation) technique has been used. The parameters that has been used to try in the 10-fold Cross Validation process were C = {1, 2, 10, 50, 100, 250, 500, 750, 1000} γ= {0.001, 0.01, 0.1, 0.5, 1, 2}. Figure 3 shows the optimal parameters of LR methods and the figure 4 shows the results that has been compared with the singular models of [5] and [4]. In the figure 4, the performance of complete (full) features is 99.9381%, but by using the feature selection method we can get the similar performance results or get a better result than when we have been using full features. So focusing on the performances we can observe that some of the features in KDDCUP'99 may have negative impact on the accuracy. The figure 5 compares the performance of TN, FN, FP and TP.



Feature Extraction for Different Models from KDD CUP'99

Total Features in KDDCUP'99	41	
DA Model	2, 12, 23, 24, 29, 31, 32, 36, 39	No. of features Used :14
LR Model	2, 12, 23, 24, 29, 31, 32, 36, 39	No. of features Used : 9
ICMP_LR Model	5, 24, 31, 32, 33, 37	No. of features Used : 6
UDP_LR Model	3, 5, 6, 8, 27, 29, 34, 35, 36, 40	No. of features Used: 10
TCP_LR Model	4, 13, 22, 24, 27, 28, 30, 32, 33, 34, 35, 36	No. of features Used:
PCA Model	2, 3, 4, 23, 24, 25, 26, 29, 30, 33, 34, 36, 38, 39	No. of features Used: 14

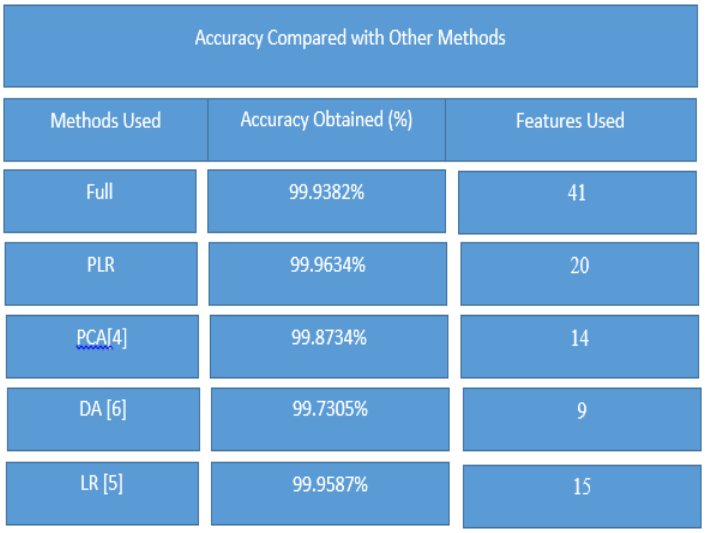
Fig -2. Feature Extraction



The Best Parameters of LR Method

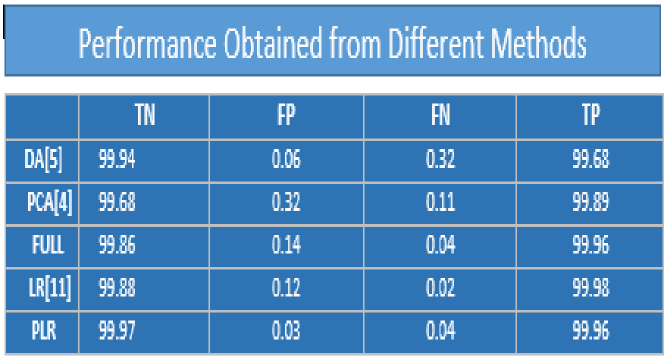
Name of Model	C	γ
Full LR	1000	0.5
TCP LR	1000	1
ICMP LR	750	2
UDP LR	2	0.1

Fig -3: The Best Parameter of LR Method



Accuracy Compared with Other Methods		
Methods Used	Accuracy Obtained (%)	Features Used
Full	99.9382%	41
PLR	99.9634%	20
PCA[4]	99.8734%	14
DA [6]	99.7305%	9
LR [5]	99.9587%	15

Fig -4: Accuracy Comparison

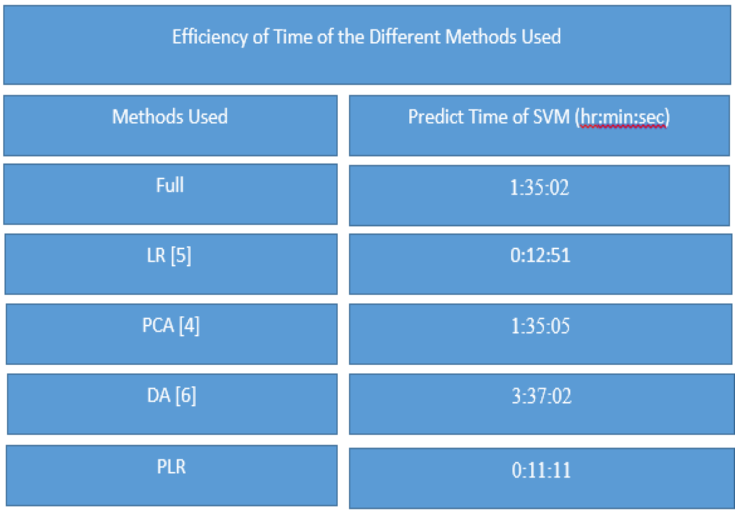


Performance Obtained from Different Methods				
	TN	FP	FN	TP
DA[5]	99.94	0.06	0.32	99.68
PCA[4]	99.68	0.32	0.11	99.89
FULL	99.86	0.14	0.04	99.96
LR[11]	99.88	0.12	0.02	99.98
PLR	99.97	0.03	0.04	99.96

Fig -5: Performance Obtained from Different Methods

According to Ambwani [2], the 99% of accuracy can also be achieved within a prediction time of 7.35 sec only, but having a number of 6890 samples only. Here in this paper study, number of full dataset that has been used is 4,898,430 (KDDCUP'99 full data) which is approx. 710 times larger than that of data used in [2]. This experiment has taken 11min and 10sec. This gives an approximation that we can save prediction by 10 times, thus it becomes a more efficient method for intrusion detection by finding accuracy.

The prediction times in the figure 6 bellow.



Efficiency of Time of the Different Methods Used	
Methods Used	Predict Time of SVM (hr:min:sec)
Full	1:35:02
LR [5]	0:12:51
PCA [4]	1:35:05
DA [6]	3:37:02
PLR	0:11:11

Fig -6: Efficiency of Time of the different Methods Used

5. CONCLUSIONS

In this work the protocol based intrusion detection system has been practiced using the data set KDDCUP'99 which contains 41 features that has been distributed to different protocols and different machine learning models like LR, PCA, LDA classification has been compared that has been shown in different table image. In this paper we have worked on an Anomaly- Based Network Intrusion Detection System using- 1.Feature Extraction, 2. Classification & 3. Recognition.

The network based intrusion detection systems which are used traditionally are slower than protocol base intrusion detection system.

Using protocols for intrusion detection will be more beneficial while using with other latest technologies like M2M communication systems and IOT concepts.

ACKNOWLEDGEMENT

We take the opportunity to express our gratitude and thanks for the excellent guidance, supervision and continued encouragement of my Guide **Prof. Hanumantha Rao** this project. We also take this opportunity to express deep gratitude to our friends for their support and inspiration that help me complete this work through its various stages. We are responsible for providing valuable information provided to faculty members of the Computer Application Department at JSSATE Noida in their respective fields. We are grateful for their cooperation during our appointment. Finally, thanks to the continued encouragement of our faculty members, this appointment would not have been possible without you.

REFERENCES

[1] Daramola O. Abosede, Adetunmbi A. Olusola, AdeolaS. Oladele,. "Analysis of KDD'99 Intrusion Detection Dataset for Selection of Relevance Features", Proceedings of the World Congress on Engineering and Computer Science, Vol. I, October 20-22, 2010

[2] Ambwani, T., 2003, "Multi class support vector machine implementation to intrusion detection," Proceedings of the International Joint Conference of Neural Networks, vol. 3, pp. 2300-2305, January 27.

[3] Pavel, L., Patrick, D., Christin, S. and R. Konrad, 2005, "Learning intrusion detection: supervised or unsupervised?" 13th International Conference on Image Analysis and Processing, pp. 50-57.

[4] Lai, C. Y., 2007, "A Novel Approach to Multi-classifier Based on Multiple feature Sets with SVM for Network Intrusion Detection," Chung Hua University thesis, July 24.

[5] Huang, W. C., 2007, "Using Regression Theiry fir Feature Selection in Intrusion Detection System," Chung Hua University thesis, August 10.

[6] Wong, W. T., and Lai, Y. C., 2006, "Identifying Important Features For Intrusion Detection Using, Discriminant Analysis And Support Vector Machine," International Conference of Machine Learning and Cybernetics, Vol. 6, pp. 3563-3567.

[7] Freemont Avenue Software, 2004, " White paper Protocol Anomaly Detection," September 9.

[8] Erwan Lemonnier – Defcom., 2001, "Protocol Anomaly Detection in Network-based IDSs", June 28.

[9] Cortes, C., and Vapnik, V., 1995, "Support-Vector networks," Machine learning, Vol. 20, pp. 273-297.

[10] Grandvalet, Y., and Canu, S., 2002, "Adaptive scaling for feature selection in SVMs," Neural Information Processing System, Vol. 15, pp. 553-560.

[11] Hsu, C. W., Chang, C. C., and Lin, J. C., 2003, "LIBSVM:a library for support vector machines," Available http://www.csie.ntu.edu.tw/~cjlin/libsv

[12] Smola, A. J., 1998, " Learning with Kernels PhD Thesis", GMD, Birlinghoven, Germany.

[13] Jolliffe, I. T., 2002, "Principle Component Analysis" Springer, 2nd ed., New York, USA.

[14] Stolfo, S. J., Wei, F., Wenke, L., Prodromidis, A., and Chan, P. K., 2000, "Cost-based modeling for fraud and intrusion detection: results from the JAM project," Proceedings of DARPA Information Survivability Conference and Exposition, vol. 2, pp. 130-144.