IJBPAS, April, 2018, 7(4): 515-526

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

NEW APPROACH FOR CLASSIFICATION R2L AND U2R ATTACKS IN INTRUSION DETECTION SYSTEM

RAFEEF FAUZI NAJIM AL-SHAMMARI

Department of Computer Science, College of Science, University of Kerbala, 56001, Karbala, Iraq

***Corresponding Author:** Rafeef Fauzi Najim Al-Shammari: E Mail: Rafeef.fauzi@uokerbala.edu.iq

Received 26th Oct. 2017; Revised 1st Dec. 2017; Accepted 29th December 2017; Available online 1st April 2018

ABSTRACT

With the development of web the world has changed into a worldwide advertisement platform with all financial and business practices being conveyed on the web. Being the most basic asset of the creating scene, it is the helpless protest and thus should be secured from the clients with perilous identity set. Since the Internet does not have central observation segment, attackers once in a while, using different progressive hacking topologies find a way to sidestep frameworks security and one such gathering of attacks is Intrusion. An intrusion is a development of breaking into the structure by trading off the security game plans of the structural set up. The strategy of taking a gander at the framework data for the possible intrusion is known to be intrusion detection. Throughout the previous two decades, programmed intrusion discovery framework has been an essential point of thorough investigation. Till now scientists have created Intrusion Detection Systems (IDS) with the ability of recognizing assaults in a few accessible situations; most recent on the scene are Machine Learning approaches. In this paper, the preprocessing unsupervised discretization and feature selection method has been applied to enhance the classification accuracy. Unsupervised discretization method is extremely important to make NSL-KDD data set as appropriate input for experiment. The discretization method needs to map these non numeric values into numeric values of features for helping classifier. After discretized the data, Principal Component Analysis (PCA) method is applied to generate subset features from whole data set. PCA method has been used to reduce dimensionality from dataset. The Naïve Bayes classifier has been employed for classification data as synonymous attacks or normal. For experimental analysis, the NSL-KDD standard data has been used to evaluate the proposed model. The empirical analysis results of proposed model demonstrate that it is better in terms of all performance measures. A comparative analysis of the results obtained for the proposed model using preprocessing methods and existing NaiveBayes algorithm with original data is presented. The empirical results prove that the performance of the proposed model is more robust and better than the performance of existing method.

Keywords; IDS, NSL-KDD, proposed model, R2L and U2R attacks

1. INTRODUCTION

In the advanced system, IDS has turned into a vital and vital piece of general security design. Keeping in mind the end goal to characterize an IDS, it is critical to comprehend "what is an intrusion and afterward "what is an intrusion detection. An intrusion can be portrayed as far as classification, uprightness, and accessibility. An occasion or activity causes rupture of secrecy in the event that it permits to get to assets, living in a PC in an unapproved way. An occasion or activity causes rupture of honesty in the event that it permits to change the conditions of assets, living in a PC in an unapproved way. So also, an occasion or activity causes rupture of accessibility on the off chance that it denies authentic clients to get to assets or administrations, living in a PC. IDS is a process used to handle the events which happened in sight of the computer system and network and filtering through them detecting the intrusion. IDS is working as software and hardware which use to monitor and analyzing the event for any network. With the quick development of intrusion, few intrusion detection system have been proposed in the writing. In spite of the fact that the proposed frameworks vary from each other in a few or many angles, there are some essential parts that are available in all the proposed frameworks. A generic architecture of a characteristic intrusion detection system is displayed in figure 1.

Figure 1 is simple essential system used to protect the network from any intrusions. The system is implemented for auditing and analyzing the event from inside and insight the network, it gives alarm when the attacks are found. The heart of the IDS is Processing unit where all research tried to apply their algorithm for improving the detection of intrusion. With growing the intelligence hackers, they have developed new method while the existing system cannot detect these types attacks. Due all the existing IDs system have big data base which store all the batch files of attacks. The attacks use method for slightly change the batches files and the existing system is unable to detect these batch files. Currently, the machine learning is used to validate the problem. The machine learning is able to detect any difference in the batch files. The main object of classification problem depends on decrease in the probability of error while the classification algorithms implemented. Consequently, the key point is the means by which to pick a classification approach to develop accurate intrusion

detection systems with respect to high detection rate whilst keeping a low false positive rate. The proposed work that combines Naïve Bayes classifier with discretization and PCA methods for enhancing classification accuracy of intrusion detection system. The User to Root (U2R) and Remote to Local (R2L) attacks have been selected from NSL-KDD data for testing the proposed model. It is observed that the proposed model has improved the classification accuracy.

The paper is organized as follows: section 1 is an introduction. The related work is introduced in section 2. In section 3 is presented the proposed model. The experimental analysis is given in section 4. Finally, the paper is closed with conclusion in section 5.

2. RELATED WORK

This section summarizes Numbers of research outcomes in relation to intrusion detection system using different machine learning algorithms. The researchers have implemented different various machine learning techniques for solving the problem. The summary of few of the existing reach work intrusion detection using machine learning is shown in table 1.

3. PROPOSED MODEL

Classification intrusion detection system is the main target of present research work. In this section is presented the formwork of proposed model for detection intrusion. Figure 2 illustrates the proposed model for detection R2L and U2L attacks. The present research work has used NSL-KDD standard intrusion detection system data set for examining the proposed model. The discretization method has been applied to find out a set of cut points, which separate the range into small number of intervals. Discretization method has helped to increase the accuracy of classification. Due to dimensionality reduction in the intrusion detection data set the PCA method has been used to select subset features from the original dataset. These subset features have been processed by Naïve bays algorithm. It is observed that the proposed model increases the accuracy and the time of classification is decreased. Finally a comparative results analysis between the proposed model with preprocessing and existing classifiers with original data is presented. The detailed description of each step used in the proposed model is presented in the subsequent subsection.

3.1 NSL-KDD DATA SET.

The NSL-KDD data is intrusion detection system data set. NSL-KDD data set

is updated version of KDD cup'99 data set. This data has been developed by McHugh [8]. NSL-KDD data set contains 4,898,431 instances. The NSL-KDD is used operation system and its application for collection of the network packets. NSL-KDD has contains 42 features and four main attack types. These four main attacks have 37 types of attacks. The NSL-KDD data set has been labeled as attacks and normal. In the present research work has used Root and Remote to Local types. Table 2 shows all types User to Root (U2R) and Remote to Local (R2L) attacks.

3.2 DISCRETIZATION METHOD

Discretization is a preprocessing technique used to process of mapping continuous features into nominal features. Discretization method employs to find out the cut points of instances values and is used to divide them into small interval values. The cut point is real numbers, it is split into two interval values one is great value and another is less than or equal cup point. The main object of used Discretization method is to determine the number of values of continuous attributes by dividing the range of the features into small intervals. Splitting the values of features into small interval helps to reduce the complexity from the original data. The discretization method is used to increase the performance of classification. In this proposed model the unsupervised method has been applied to improve the NSL-KDD data set before applying classifiers.

3.3 FEATURE SELECTION METHOD

Feature selection is an extremely important phase in intrusion detection system due to the fact that the network data is more complex and is noisier in specific behavior. Furthermore, the feature selection methods are necessary for reducing dimensionality from dataset. When dimensionality is reduced from data set the classifier obtains the highest performance. In the present research work the Principal Component Analysis has applied for selection subset feature from original data set. The PCA has selected the features which are been more significant among whole of the features. The detailed description of PCA method is as follows:

A. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a very important method for dimensionality reduction of a data set. If the data set consists of a large number of redundant irrelevant features. PCA is worked by transforming the data set to a new sub set of ordered variables so that the first few variables retain most of the variations in all of the original variables. The PCA method is used for the least square decomposition to

convert linear projection of multivariate high dimensional data onto low-dimensional subspace. Principal component analysis targets to find out the orthogonal directions of strong variability in data. Given a set of observed d-dimensional independent data vectors x_i where $i \in 1,\cdots,n$ the orthogonal projection is executed by

$$\boldsymbol{Y_t = A^T(x_t - \mu)}\text{……(1)}$$

The transformed data is y where μ is the sample mean of the observe data. The sample covariance matrix is as follows:

$$S = \frac{1}{n}\sum i = 1^n_(x_i - \mu)(x_i - \mu)^T \text{……..(2)}$$

The reconstruction error is computed by E_r:

$$E_r = \left\| x - \mu - y_{A_q} \right\|^2 \text{……(3)}$$

When applying the PCA, 7 features are generated that are most significant out of 41 original features which are listed in table 3.

3.4 NAÏVE BAYES ALGORITHM

Bayesian classification is called nave as the computation involved in it is very simple. Nave Bayesian algorithm is used to classifies intrusion detection system as attacks or normal. The Bayesian algorithm is depended on and the probability of the event that will be occur in future with use of the previous occurring of the same event. Bayesian algorithm is working as follows: Suppose $T = T_1, T_2, T_3, \cdots T_n$ is as feature vector of intrusion detection system features and the values of the features are $T_1, T_2, T_3 \cdots, T_n$ and also is number of features in the NSL-KDD dataset. The Y indicates the class of IDS data as Attacks and Normal. The Bayes equation is

$$\boldsymbol{P(Y|T) = \frac{P(Y).P(T|Y)}{P(T)}} \text{……..(4)}$$

In Equation-6.2, the a-priori probability is $P(T)$ of a randomly picked intrusion detection system feature values. Where Y, and $P(Y|T)$ are a-prior probability of class. The Y has T indicates the probability of a randomly picked IDS with class Y.

3.5 PERFORMANCE MEASURES METHODS

The performance measures have been conducted to examine the results of proposed model. The Accuracy, False Positive, F-measure, Precision, True Positive and Time are employed. Table 4 displays a simple confusion matrix for two prediction intrusion classes.

The performance measures equations are as follow:

$$\textbf{False Positive Rate (FPR)} = \frac{FP}{TN+Fp}\% \textbf{ 100… (5)}$$

$$\textbf{True Positive Rate (TPR)} = \frac{TP}{TP+FN}\%\textbf{100…… (6)}$$

$$\textbf{Accuracy}= \frac{TP+TN}{TP+FP+TN+FN}\%\textbf{100…… (7)}$$

$$\textbf{Precision}=\frac{TP}{TP+FP}\%\textbf{100…….. (8)}$$

$$\boldsymbol{F-measure} = \frac{2*Precision*Recall}{Precision+Recall}\ \%\textbf{100….(9)}$$

4. EXPERIMENTAL SETUP

The proposed model intrusion detection system method is implementing by using

Python with 64 windows 10 Ultimate with the core i5 processor and 4 GB RAM. Different various evaluation metrics have been used to test the proposed model. In this experiment proposed model of Naïve Bays, discretization and PCA algorithms are applied. In this experiment, two types of attacks R2L and U2L are selected. The data has only 6097 attacks and normal instances are used. Table 4 shows the performance of the proposed model. It is observed that the correct classification of instance is 4706 out 1391 instances.

To improve the proposed model, it a decision is made to use preprocessing. discretization and PCA methods . The unsupervised discretization method has been used for mapping numerical-valued features in data set to nominal-valued features for enhancing the proposed model. Discretization method applied to make the NSL-KDD data set values in specific range for increasing the performance and reducing the time of classification. After discretization method is applied the PCA feature selection method is employed for dimensionality reduction. Seven features have been selecting from original data which has more ranking. Theses subset features have been used for classification by using Naïve Bayes algorithm for detecting the intrusion. Table 6 shows the performance of Naïve Bayes classifier with original data. The weighted averages of different performance measures are very less. It is necessary to improve the results of classifier with original data. The preprocessing methods have been applied to improve the results of existing classifier with original data. The results of proposed model are shown in table 7. It is observed that the proposed model has outperformed with preprocessing method better than the original data.

The TP, FP, Precision, Recall, F-Measure and accuracy have been used to test the proposed model. These performance measures are used to compare the proposed model with different existing model. The results of proposed model are compared with Naïve bays classifiers with original data. The tables 6, 7 summarize the results of the proposed model with preprocessing method and Naïve nays classifier with original data. It is observed that the proposed model has outperformed. Figure 3 displays accuracy results performance of the proposed model in comparison with Naïve nays classifiers with original data. Finally, it is concluded that the proposed model can detect different types of attacks with best accuracy compared to other existing systems.

Table 1: Intrusion detection system techniques

Authors	Preprocessing Techniques	Algorithms	Data set	Methodology	Performance measures
Jiong Zhang et (1)	Extracting the attributes from data set by using by feature selection method	Random forests algorithm	KDD cup data set	For developing intrusion detection	Accuracy, FP
Xin Du, Yingjie et (2)	Information entropy method	K-means algorithm	MIB network	For developing intrusion detection	Accuracy
Gharehchopogh FS. et al. (3)	They divide data set into four categories namely fundamental, host, network and based on time attributes methods	K-means and Fuzzy K-means algorithms	KDD cup data set	For detecting DoS attack	DR, Accuracy
Miller, W.Deritrik et al. (4)	Histogram payload packets method	Den Stream and frequency histogram algorithms	DARPA and MCPAD data sets	For developing intrusion detection	Accuracy, DR, FP
Ching Chen R et al. (5)	Rough Set Theory method	Support Vector Machine (SVM) algorithm	DARPA dataset	For developing intrusion detection	DR, Accuracy
Dubey GP et al. (6)	Rough Set Theory method	Increment SVM algorithm	DARPA data set	For developing intrusion detection	Accuracy
Shrivastava SK et al. 2011 (7)	Rough set theory method	Support vector machine algorithm	KDD cup data set	Intrusion detection	Accuracy, FP
lakhina S et al. (8)	PCA method	Neural network algorithm	DARPA data set	For developing Intrusion detection system	Accuracy
Lath R. et al. (9)	Statistical normalization method	K-means, SVM, K-nearest Neigh bour algorithms	DARPA data set	For developing Intrusion detection	FP, Accuracy
Neethu (10)	PCA method	Naïve Bayes algorithm	KDD cup data set	Reduce false positive for intrusion detection	Accuracy, DR, FP
Jain YK et al. (11)	Information entropy method	NB and Bayes Net algorithms	KDD cup data set	For increase detection rate of intrusion detection	Accuracy, Recall, Precision, F-Measure
Al-Nafjan K et al. (12)	PCA method	Modular neural network algorithm	KDD cup data set	For reduce false positive rate of intrusion detection	RMSE
Yingjie Zhou et al. (13)	Not mentioned	Graph Mining algorithm	Abilene data set	For developing Intrusion detection	Accuracy
Tian X, et al. (14)	Feature selection method	Fast Decision Tree algorithm	peer-peer Bit Torrent, PP Live data set	For classification Stream network traffic	Accuracy, DR

Table 2: All types of U2R and R2L attacks in NSL-KDD

Attacks in Dataset	Type of attacks
R2L	Guess_password, Ftp_write,Imap, Phf, Multihop, Warezmaster Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named
U2R	Buffer_ overflow, Loadmodule Rootkit, Perl,Sqlattack, Xterm, Ps

Table 3: Seven significant attributes obtained by PCA method

Feature set	Two rank for all 7 features
service src_bytes dst_bytes srv_count dst_host_same_srv_rate dst_host_srv_rerror_rate	0.9600 0.9258

Table 4: Confusion matrix

	Prediction Intrusion class		
Actual Class	Activity	Attack	Normal
	Attack	TP	FN
	Normal	FP	TN

Table 5: Performance analysis of proposed model

Performance	Proposed model
Time	0 seconds
Correctly Classified Instances	4706
Incorrectly Classified Instances	1391
Total Number of Instances	6097

Table 6: Results of Naïve Bayes classifier with original data

Weighted Avg.	TP Rate	FP Rate	Precision	Recall	F-Measure	Total accuracy	Classes
	0.857	0.001	0.998	0.857	0.922		guess_passwd
	0.184	0.003	0.400	0.184	0.252		buffer_overflow
	0.704	0.004	0.988	0.704	0.822		warezmaster
	0.427	0.011	0.714	0.427	0.534		snmpgetattack
	0.962	0.006	0.883	0.962	0.921		httptunnel
	0.633	0.020	0.138	0.633	0.226		ps
	0.970	0.045	0.722	0.970	0.828		snmpguess
	0.089	0.006	0.093	0.089	0.091		multihop
	0.588	0.005	0.400	0.588	0.476		named
	0.286	0.025	0.050	0.286	0.085		sendmail
	0.429	0.003	0.240	0.429	0.308		loadmodule
	0.615	0.003	0.457	0.615	0.525	77.18	xterm
	0.050	0.001	0.250	0.050	0.083		rootkit
	0.500	0.004	0.250	0.500	0.333		xlock
	0.714	0.000	1.000	0.714	0.833		perl
	1.000	0.013	0.092	1.000	0.168		xsnoop
	1.000	0.000	1.000	1.000	1.000		sqlattack
	0.533	0.078	0.016	0.533	0.032		ftp_write
	0.722	0.002	0.481	0.722	0.578		imap
Weighted Avg.	0.800	0.004	0.242	0.800	0.372		imap

Table 7: Results of proposed model

Weighted Avg	TP Rate	FP Rate	Precision	Recall	F-Measure	Total accuracy	Classes
	0.997	0.001	0.999	0.997	0.998		guess_passwd
	0.934	0.008	0.587	0.934	0.721		buffer_overflow
	0.994	0.007	0.985	0.994	0.990		warezmaster
	1.000	0.000	1.000	1.000	1.000		snmpgetattack
	0.977	0.001	0.981	0.977	0.979		httptunnel
	0.600	0.002	0.545	0.600	0.571		ps
	1.000	0.000	0.997	1.000	0.998		snmpguess
	0.578	0.002	0.667	0.578	0.619		multihop
	0.559	0.001	0.826	0.559	0.667		named
	0.929	0.000	1.000	0.929	0.963		sendmail
	0.500	0.000	1.000	0.500	0.667	97.53	loadmodule
	0.385	0.001	0.588	0.385	0.465		xterm
	0.575	0.003	0.548	0.575	0.561		rootkit
	0.889	0.000	0.889	0.889	0.889		xlock
	0.857	0.000	0.857	0.857	0.857		perl
	0.000	0.000	0.000	0.000	0.000		xsnoop
	0.000	0.000	0.000	0.000	0.000		sqlattack
	0.067	0.000	1.000	0.067	0.125		ftp_write
	1.000	0.000	0.900	1.000	0.947		imap
	0.800	0.000	1.000	0.800	0.800		imap
Weighted Avg	0.975	0.003	0.976	0.975	0.974		

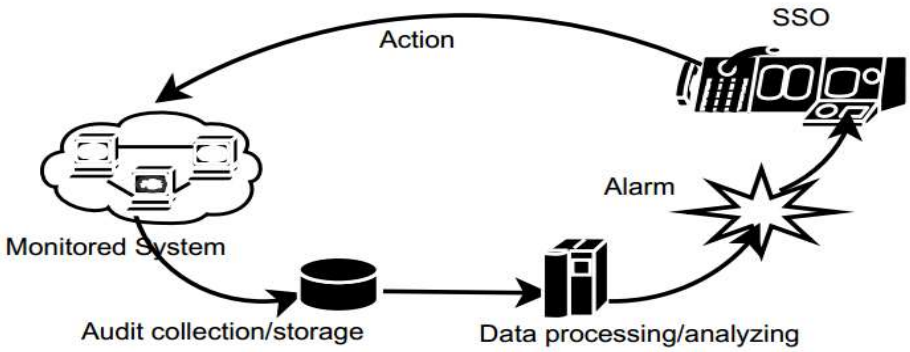

Figure 1: A generic architecture of IDS

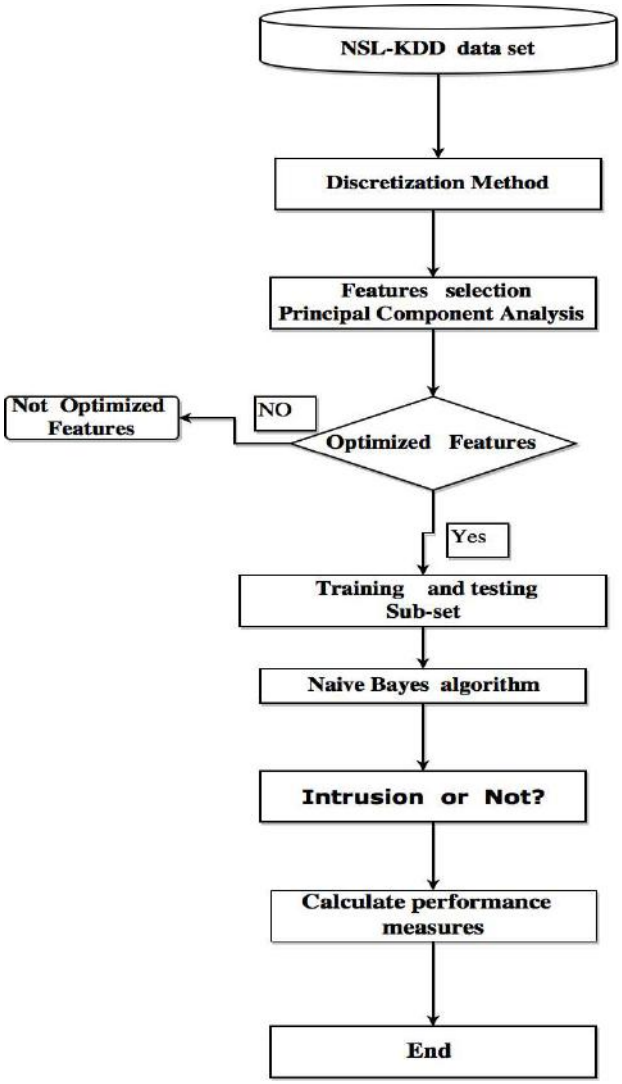

Figure 2) Proposed model

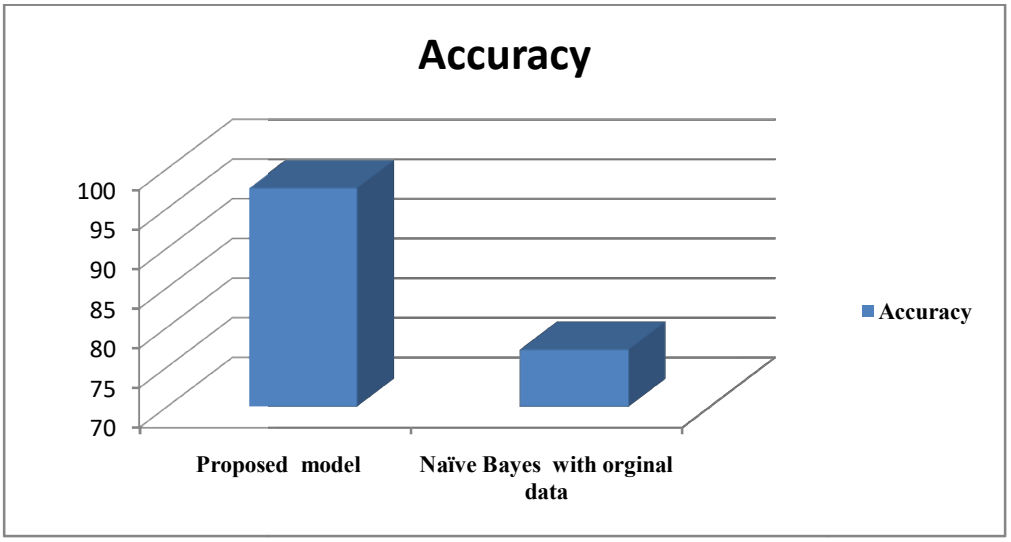

Figure 2: Performance accuracy of proposed model and existing model.

5. CONCLUSION

The main target of present study is to enhance the existing Naïve bays algorithm for developing robust intrusion detection system. In this work the proposed model of combined preprocessing techniques is presented. The discretization method has been applied to range the values of features in data set. In this research U2L and R2L attacks have been used. However, the PCA features selection method is applied to solve the dimensionality reduction problem.. The PCA method has selected 7 most significant subset features from the original 41 features. The proposed model has used subset features for classification intrusion detection system. The different and various f evaluation metrics are applied to test the proposed model. It is observed that the accuracy of proposed model is very high in comparison with the existing model. The accuracy of proposed model is 97.33%. A comparative results analysis between the proposed model and existing naive bays model is presented. It is concluded that the proposed model has outperformed the naïve bays with original data. In future, the researcher will try to use different types of attacks.

6. REFERENCES

[1] J. Zhang, M. Z. Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection. 06. IEEE International Conference on communication 2006. DOI: 10.1109/ICC.2006.255127

[2] Xin Du, Yingjie Y, Xiaowen K. Research of Applying Information Entropy and Clustering Technique on Network Traffic Analysis. CIS '08. International Conference on Computational

Intelligence and Security 2008. DOI: 10.1109/CIS.2008.132

[3] Gharehchopogh FS, Jabbari N, Azar ZG. Evaluation of Fuzzy K-Means and K-Means Clustering Algorithms in Intrusion Detection Systems. International journal of scientific and technology research, 2012, 11(1), 66-71.

[4] Miller Z, Deitrick W, Hu W. Anomalous Network Packet Detection Using Data Stream Mining. Journal of Information Security, 2011 (2), 158-168. doi:10.4236/jis.2011.24016

[5] Ching Chen R, Cheng K and Hsieh C. Using Rough set and Support Vector Machine for intrusion detection. International Journal of Network Security & Its Applications (IJNSA), 2009. 1(1),1-12

[6] Dubey GP, Gupta N. Bhujade RK. A Novel Approach to Intrusion Detection System using Rough Set Theory and Incremental SVM. International Journal of Soft Computing and Engineering (IJSCE), 2011), 1(1) 2231-2307

[7] Shrivastava SK and Jain P. Effective Anomaly based Intrusion Detection using Rough Set Theory and Support Vector Machine. International Journal of Computer Applications, 2011 3(18), 0975 – 8887

[8] lakhina S., Joseph S and Verma B. Feature Reduction using Principal Component Analysis for Effective Anomaly–Based Intrusion Detection on NSL-KDD. International Journal of Engineering Science and Technology, 2010, 2(6), 1790-1799.

[9] Lath R, Shrivastava M. Analytical Study of Different Classification Technique for KDD Cup Data .2011 99(3)

[10] Neethu B. Classification of Intrusion Detection Dataset using machine learning Approaches. International Journal of Electronics and Computer Science Engineering, 2012, 2277-1956/V1N3-1044-1051

[11] Jain YK, Upendra, Intrusion Detection using Supervised Learning with Feature Set Reduction. International Journal of Computer Applications, 2011 33(6), 0975 – 8887

[12] Al-Nafjan K, Al-Hussein MA., Alghamdi AS. Haque MA and Ahmad I. Intrusion Detection Using PCA Based Modular Neural Network. International Journal of Machine Learning and Computing, 2012. 5(2), 583-586

[13] Zhou Y, Guangmin, H, He W. Using graph to detect network traffic anomaly. IEEE International Conference on Communications, Circuits and Systems,

2009.
DOI: 10.1109/ICCCAS.2009.5250514

[14] Tiani X, Sun Q, Hunga X, Yan M. A Dynamic Online Traffic Classification using Data Stream Mining. IEEE, Multimedia and Information Technology, 2008. DOI: 10.1109/MMIT.2008.185