Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review Paper on Data Mining Approach for Feature Selection for Network Intrusion Detection System

Tanya Garg, Amandeep Kaur

M.Tech (C.S.T) Central University of Punjab,
Bathinda , India

Abstract— Intrusion Detection Systems in Computer Network detects intrusion on the basis of some attributes of the Network Traffic. These attributes provide a way to detect Intrusion activity from the coming Network Traffic. To detect these intrusions, basically Training dataset is to be used i.e. KDDCup'99 is used for identifying intrusions in the Network Traffic by the Network Intrusion Detection System. In this paper I am going to discuss the overall mechanism used to reduce the feature set i.e. KDDCup'99 IDS Dataset having 41 features to reduce the computational time and training time of Network Intrusion Detection System.

Keywords— Attack Categories,KDDCup'99 Dataset, NIDS Models, Techniques for Feature Reduction, Classifiers, Performance Metrics, NIDS (Network Intrusion Detection System) Datasets

I. INTRODUCTION

In Daily life, Networks and the computers get more complex day by day. This means that there are also more and more services available for malicious exploitation. New vulnerabilities are found from common programs daily and even on vulnerability in a single computer might compromise the network of an entire company. There are two parallel ways to address this threat. The first way is to ensure that a computer doesn't have any known security vulnerabilities, before allowing it to the network it has access rights. The other way, is an Intrusion Detection System. IDSs concentrate on detecting malicious network traffic, such as packets that would exploit a known security vulnerability.

According to [1], a computer network intrusion is defined as any set of actions that attempt to compromise the integrity, confidentiality or availability of a network resource. In general, intrusion attempts are external malicious actions that have the purpose of intentionally violating the system security properties. An intrusion detection system (IDS) monitors network traffic and monitors for suspicious activity and alerts the system or network administrator. In some cases the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user or source IP address from accessing the network .In complex domains, such as Network Intrusion Detection System (NIDS), a huge amount of activity data is collected from the network generating large log files and raw network traffic data, in which human inspection is impossible. Thus, these activity data must be compressed into high-level events, called attributes. After it, a set of attributes is obtained and monitored by the NIDS in order to detect intrusion attempts. However, there are some attributes with false correlations, hiding the underlying process, and other that may be either irrelevant or redundant (its information is somehow included in other attributes). In this way, removing these attributes, or rather, selecting an optimal attributes set that adequately describes the network environment are essential in order to achieve fast and effective response against attack attempts, reduce the complexity and the computation time, and increase the precision of the NIDS. In this way, development of methods for selecting optimal attributes is required.

II. DATA MINING

A Machine Learning Approach, Data mining is a part of KDD (Knowledge Discovery in Database). It is a process of transforming raw data in useful information. "Data Mining is the process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions" [18]. Simoudis 1996.Data Mining is a process that analyzes huge amount of data and predicts useful information or data from it [3].It is a process of discovering patterns in data [4].

II.I. Data Mining Tasks

It basically performs two tasks:

Predictive Task: The task that predicts the value of attributes based upon the value of other attributes.

Descriptive Task: The task that is used to derive patterns that shows the relationships among data.

II.II. Dimensionality Reduction

It is an effective approach to reduce the large amount of data.

There are two approaches for Dimensionality Reduction:

Feature Selection: It is a technique to select most relevant feature set by eliminating irrelevant and redundant features. Only a subset of original features are selected

Feature Reduction: It is a process of using all the original features to make linear combinations to form a reduced set of features.

III. FEATURE SELECTION

A process that chooses an optimal subset of features according to a particular objective function.
Objectives:

- ➢ To reduce dimensionality and to remove noise.
- ➢ To improve mining performance i.e. to improve
- Speed of learning
- Predictive Accuracy
- Simplicity and Comprehensibility of Mined Results.

IV. DATA SETS

The KDDcup99 Intrusion Detection Datasets are based on the 1998 DARPA Dataset which provides a benchmark for data miners to reduce features for an Intrusion Detection Model [6].DARPA 1998 Dataset consists of about 5 million records having attacks falling into four categories and fifth for normal traffic on the network. These records were created by analysing seven weeks of network traffic on LAN environment created for military security activity. This network traffic was considered to have 41 features. On the basis of these 41 features/attributes of network traffic intrusions were detected and reported. But this dataset was highly overloaded and drop packets on the way and there was no mechanism to check the possibility of these dropped packets so, a new dataset called KDDcup99 was introduced which was made by analysing two week s traffic on the network [6]. KDDcup99 dataset consists of 41 features and each instance in record is labelled as either normal or an attack. In this database, there are 4.94,021 instances in which 92,278 are normal and 396,744 are labelled as attacks that are of 22 types and can be classified in four categories [4]:

- ➢ **Denial of Service (DoS):** An attack belonging to this category tries to make requested web server unavailable to the users.
- ➢ **Probing:** attacks that scan the entire network to get the secret information on the network and sends huge amount of packets to a number of hosts in a short time with very short duration.
- ➢ **Remote to Local (R2L):** attacks in which an unauthorized user on the entire network tries to access the account of an authorized user on the network to gain access of information of a particular host on a network.

User to Root (U2R): attacks in which an authorized user of a network tries to get an access of information as a root user.

NSL-KDD Dataset: It is the reduced KDDcup 99 dataset that has no redundant and duplicate record in its database. For our Feature Selection process, any dataset can be used according to hardware configuration of our system. One can also select random records or instances from these datasets to make a new reduced dataset to be used for their experiments.

TABLE I
NIDS ATTRIBUTES/FEATURES [4]

Sr. No.	Attribute	Sr. No.	Attribute
1.	Duration	22.	is_guest_login
2.	Protocol_type	23.	count
3.	Service	24.	srv_count
4.	Flag	25.	serror_rate
5.	Src_bytes	26.	rerror_rate
6.	Dst_bytes	27.	Srv_rerror_rate
7.	land	28.	Same_srv_rate
8.	Wrong_fragment	29.	diff_srv_rate
9.	urgent	30.	Srv_diff_host_rate
10.	hot	31.	Dst_host_count
11.	Num_failed_logins	32.	Dst_host_srv_count
12.	Logged_in	33.	Dst_host_same_srv_rate
13.	Num_compromised	34.	Dst_host_diff_srv_rate
14.	Root_shell	35.	Dst_host_srv_rerror_rate
15.	Su_attempted	36.	Dst_host_rerror_rate
16.	Num_root	37.	Dst_host_srv_serror_rate
17.	Num_file_creations	38.	Dst_host_serror_rate
18.	Num_shells	39.	Dst_host_srv_diff_host_rate
19.	Num_access_files	40.	Dst_host_same_src_port_rate
20.	Num_outbound_cmds	41.	Dst_host_srv_host_rate
21.	Is_host_login		

© 2013, IJARCSSE All Rights Reserved

TABLE 2
ATTACKS PER CATEGORY [17]

DOS	PROBING	R2L	U2R

back(1026)
land(11)
Neptune(10401)
Pod(69)
Smurf(7669)
Teardrop(15)
Normal(2573)
	Ipsweep(586)
	Nmap(151)
	Portsweep(155)
	Satan(16)
	Normal(1704)
		ftp_write(8)
		guess-passwd(53)
		imap(11)
		multihop(11)
		phf(5)
		spy(4)
		warzclient(60)
		warezmaster(20)
		normal(1934)
			Loadmodule(10)
			Rootkit(7)
			Perl(3)
			Normal(1676)
			Buffer-overflow(21)

V. SIMULATION TOOLS

Most widely used simulation environments for Feature Selection are as below:

- WEKA (Waikato Environment for Knowledge Analysis)
- RAPID MINING TOOL
- MADAM ID (Mining Audit Data for Automated Model for ID)
- ADAM (Audit Data Analysis and Mining)
- STATISTICA MINING TOOL
- KEEL (Knowledge Extraction Based on Evolutionary Learning)
- MATLAB DATA MINING SOFTWARE

VI. FEATURE SELECTION TECHNIQUES

V.I.Filter Model: It is that feature selection model which filters the data set without using any learning algorithm by just determining the internal characteristics of the data set [4].The techniques used by filter model for optimal attribute selections are based on ranking search method i.e. ranker search method is used along with single attribute evaluators to rank the importance of features and by eliminating the attributes of higher rank can give the reduced set of features.

TABLE 3
RANKING BASED FEATURE SELECTION TECHNIQUES [17]

	Name of Technique
SINGLE ATTRIBUTE EVALUATOR	➢ Chi-squared ➢ Gain Ratio ➢ Info-Gain ➢ Principal Component ➢ Relief ➢ SVM ➢ Symmetrical Uncertain

V.II.Wrapper Model: It is that feature selection model which uses learning algorithm to evaluate the performance of each selected attribute.

TABLE 4: SUBSET EVALUATOR TECHNIQUES [17]

ATTRIBUTE SUBSET EVALUATOR	➢ Cfs Subset Evaluator ➢ Classifier Subset Evaluator ➢ Wrapper Subset Evaluator ➢ Consistency Subset Evaluator

© 2013, IJARCSSE All Rights Reserved

TABLE 5
SEARCH METHODS TO BE USED WITH SUBSET EVALUATORS [17]

SEARCH METHOD	➢ Best First ➢ Exhaustive Search ➢ Genetic Search ➢ Greedy Stepwise ➢ Random Search ➢ Race Search ➢ Rank Search

V.III. Hybrid Model: It is that feature selection model which uses both the Filter Model and Wrapper Model for Feature Selection for an IDS model.

VII. CLASSIFYING ALGORITHMS

There are several machine learning algorithms that evaluates the performance of Features and classifies the network traffic as normal or attack category. These algorithms evaluates the performance of Intrusion Detection Model based on several performance metrics. There are eight categories of Classifying algorithms to design an Intrusion Detection Model with reduced number of Features. Basically the given dataset is to be divided into two datasets one is:

Training Dataset and other is **Testing Dataset.**

Training dataset: It is a dataset that is used to build a classifier which is a process of learning something from instances in order to predict the class attributes of new coming unknown instances.

Testing Dataset: It is a dataset that is used to evaluate the learning algorithm. It is used to validate the learning algorithm. These learning algorithms when applied to training dataset makes that dataset learn to work according to that algorithm and then to check whether that algorithm is appropriate to be used with that dataset is validated by evaluated that dataset with the testing dataset. The learning algorithms are responsible for predicting the class of new unknown instances. These algorithms varies depending on the simulation tool used i.e. machine learning algorithms varies from one simulation environment to another.

VIII. PERFORMANCE METRICS

Confusion Matrix: Contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. To evaluate our system, besides the classical accuracy measure, the three standard metrics of detection rate (DR) /True Positive Rate (TPR), false positive rate (FPR) & F-measure developed for network intrusions will be used. Table 6 shows these standard metrics

TABLE 6
CONFUSION MATRIX

		Predicted	
		Normal	**Attack**
Actual	**Normal**	**TP**	**FN**
	Attack	**FP**	**TN**

Here is a confusion matrix that summarizes the number of instances predicted correctly or incorrectly by a classification model.

False positive (FP): Or false alarm, Corresponds to the number of detected attacks but it is in fact normal.

False negative (FN): Corresponds to the number of detected normal instances but it is actually attacked, in other words these attacks are the target of intrusion detection systems.

True positive (TP): Corresponds to the number of detected attacks and it is in fact attack.

True negative (TN): Corresponds to the number of detected normal instances and it is actually normal.

.**Detection Rate:**

$$DR = \frac{TP}{TP + FN}$$

False Positive Rate:

$$FPR = \frac{FP}{FP + TN}$$

F-Measure: $$F - measure = \frac{2 * TP}{(2 * TP) + FP + FN}$$

Kappa: It is the measure of agreement or disagreement between two classifiers. Its value ranges from o to 1 .0 means totally disagreement and 1 means full agreement.

ROC (Receiver Operating Characteristics): Higher the value of ROC Curve better is the performance of the classifier. These metrics are important because they measure the percentage of intrusions the system is able to detect and how many misclassifications it makes. To visualize the trade-off between the false positive and the detection rates, the ROC (Receiving Operating Characteristic) curves are also depicted. Furthermore, to compare classifiers it is common to compute the area under the ROC curve, denoted as AUC. The higher the area, the better is the average performance of the classifier.

IX. BASIC METHODOLOGY FOR FEATURE SELECTION PROCESS

- Choose appropriate dataset for Intrusion Detection.
- Split the dataset into two datasets one is: Training Dataset and other: Testing Dataset.
- Convert into suitable file format according to valid data extension that your simulation tool supports.
- Install the simulation tool to be used for Feature Selection Process.
- Choose the Feature Selection model you want to evaluate i.e. either Filter Model, Wrapper Model or Hybrid Model.
- Choose the Feature Selection Techniques according to Feature Selection Model i.e. either ranking based techniques or Subset evaluator techniques along with search methods or both the techniques.
- Evaluate the Feature selection techniques with the learning algorithms that are classifiers.
- Evaluate the performance of each learning algorithm on the basis of performance metrics.
- Propose a model giving which feature selection technique with which classifier gives more accurate results with reduced number of features.

X. RELATED WORK

During the last decade, a considerable number of studies have been carried out regarding a good feature selection policy which can choose significant features from KDD CUP99 TCPDUMP.

Chebrolu et al. [5] investigated performance of two Hybrid based Feature Selection methods Markov Blanket Model and Decision Tree and proposed Hybrid approach for Feature reduction for real data feature set having 41 attributes was reduced to Feature set having 12 attributes ,17 attributes and 19 attributes using. Performance of two feature selection methods was evaluated using two classifying algorithms i.e. Bayesian Network (BNs) and Classification and Regression Trees and then performance was evaluated by ensembling of both the algorithms on 1998 DARPA intrusion detection data set. This approach showed accuracy only to detect Normal, Dos and Probe categories but not suitable for U2R and R2L categories which are the most serious category of attacks.

A. Sung and S.Mukkamala [2] proposed an approach for Computational Intelligence (CI) using Ranking methods. They used two classes of learning machines for IDS: Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) & show that SVMs are superior to ANNs in three critical respects of IDS: SVMs train and run an order of magnitude faster; SVMs scale much better; and SVMs give higher classification accuracy. Two methods for feature ranking are presented: the first one is independent of the modelling tool Performance-Based Ranking Method (PBRM), while the second method is specific to SVMs i.e. Support Vector Decision Function Ranking Method (SVDFRM).By using this approach it conclude SVMs outperform ANNs in the important respects of

A. scalability (SVMs can train with a larger number of patterns, while ANNs would take a long time to train or fail to converge at all when the number of patterns gets large)

B. training time and running time (SVMs run an order of magnitude faster)

C. prediction accuracy.SVMs easily achieve high detection accuracy (higher than 99%) for each of the 5 classes of data, regardless of whether all 41 features are used, only the important features for each class are used, or the union of all important features for all classes are used.

Mukkamala et al. [3] investigated performance of different feature selection techniques i.e. statistical methods along with comparative study of feature ranking methods for intrusion detection systems. Feature ranking was done by chi-square analysis and logistic regression, single attribute evaluators and it was predicted that chi-square analysis produced more consistent results as compared to logistic regression. Performance of reduced dataset was evaluated by Artificial Neural Networks Classifying Algorithm. The another feature ranking methods: Forward Selection and Backward propagation also produced consistent results All the experiments were conducted on DARPA IDS Dataset in Weka Environment. Features got highly overlapped for all the attack categories when ranked using chi-square analysis. Testing and training time gradually decreased for all the categories of attacks but the accuracy increased slightly for Normal, DOS, Probe attacks but there is no accuracy for the most serious attack categories: U2R and R2L.

L. I. Kuncheva et al. [9] proposed an approach of Random Subspace method. Choosing random subsets of features is termed the random subspace method. Each classifier in the ensemble is built upon a randomly chosen subset of features of predefined size d. They suggests that good results are obtained for tree classifiers built upon d _ n/2 features, where n is the total number of features. The random subspace method has been found to work well when there is redundant information which is “dispersed” across all the features rather than concentrated in a subset of them.

Ming Yang Su et al. [14] proposed an approach on a genetic algorithm combined with kNN (k-Nearest Neighbour) for feature weighting. They proposed NIDS is described by three parts. All features that are considered in the paper, state the encoding of a chromosome and the fitness function, and finally describe the details about selection, crossover and mutation. A method to weight features of DoS/DDoS attacks, and analysed the relationship between detection performance and number of features. Different to previous works, their method can be and really have been implemented to a real-time NIDS. This is because all features applied in the paper are directly collected from packet headers. They weighed all initial 35 features in the training phase and then select features having higher weight to implement a NIDS for testing. Many DoS/DDoS attacks are applied to evaluate the system. For known attacks they got the best 97.42% overall accuracy rate while only the top 19 features are considered; as for unknown attacks, they got the best 78% overall accuracy rate by top 28 features. **Khor et al. [8]** investigated performance of two Filter Based feature selection methods CFSE (Correlation based feature selection Evaluator) and CSE (Consistency subset Evaluator) and five feature sets were developed and tested individually by conducting t-sample test. They evaluated 1998 DARPA intrusion detection dataset for performance evaluation. Shared Feature Set having minimum number of features was formed by extracting features from the two feature sets that were obtained using CFSE and CSE methods. Then the two feature sets were combined without repeating the same features and a new Combined Feature Set was generated and then Fifth Feature set was generated by adding neglected features to the Shared Feature Set and this set is the Proposed Feature set by researcher. Then they compared efficiency of all these four feature sets on the basis of accuracy with his proposed feature set using BNs Classifying Algorithm and it was found that there was no significant difference in accuracy obtained between the proposed feature sets and other feature sets except shared feature set which gave poor accuracy. They concluded that all the feature sets have comparable performance with the proposed method. Moreover the number of features in the proposed feature set is the least as compared to CFSE, CSE and Combined feature sets. The training data used was few for certain category which is Limitation of this approach. The results could be more accurate if experts would concentrate more on Feature Selection Process than on Feature Selection Methods. Performance could be better if multiple classifiers would be used.

M. Revathi and T. Rames [10] investigated the performance of filter based feature selection method i.e. Best First Search and two reduced datasets were generated by analysis on KDD Cup99 IDS Datasets. The performance of reduced datasets was evaluated using two classifying algorithms: ID3 (Iterative Dichotomiser) and J48. Simulation was performed in WEKA Environment. Performance was compared for the three different feature sets using three performance metrics that are Accuracy, Sensitivity and Specificity. Classifier Performance was evaluated under ROC Curve for both the classifying algorithms used. From the result it was concluded that ID3 is much more accurate and efficient than J48. Analysis was not done for all the five classes of attacks.

Christiane F. L. Lima et.al [4] investigated some attribute selection approaches through a comprehensive comparison of C4.5 decision-tree model based on Shannon entropy with other three attribute selection method, namely C4.5 based on Rényi entropy, C4.5 based on Tsallis entropy and an approach that combines Shannon, Rényi and Tsallis entropies. In order to obtain optimal attribute subsets that increase the detection capability of classifying network traffic as either normal or suspicious, the filter model based attribute selection technique gain ratio was considered, taking account Rényi entropy *versus* Shannon entropy and Tsallis entropy versus Shannon entropy. To evaluate the classification performance of these methods, it was considered four attack categories (DoS, Probing, R2L and U2R) based on KDD Cup 1999 dataset, and the following classification models: Clonal selection Algorithm (CLONALG), Clonal Selection Classification Algorithm (CSCA) and Artificial Immune Recognition System (AIRS).Finally the attribute selection has decreased significantly the number of attributes and data dimensionality, leading to a better performance of the AIS algorithm, resulting in lesser running time compared to the situation when the complete attributes set of the original database was used but the selected set of attributes is not performing well for all attacks categories.

Sangatanee et al. [13] investigated the performance of Filter based Feature Selection method i.e. Information Gain and proposed the Network Based On-line Real Time Intrusion Detection System by reducing Features to 12 for DOS and Probe Attacks. Three experiments were performed for evaluation of proposed Real time NIDS The proposed NIDS model was developed in WEKA tool version 3.6.0.The experiments were conducted on RLD09 (Reliability Lab Data 2009) dataset and its performance was also compared with results of evaluation on KDD Cup 99 Dataset. The selected Feature set was evaluated using several Classifying Algorithms: Decision Tree, Ripper Rule, Back-Propagation Neural Network, Radial Basis Function Neural Network, Bayesian Network, and Naive Bayesian for designing the intrusion detection system. The proposed model performed best when it was evaluated using Decision Tree Classifying Algorithm. They also proposed a Post Process technique for reducing False alarm rates. But the data set was not reduced for all type of attack categories.

Kira et al. [8] proposed filter approach is the RELIEF algorithm. The RELIEF algorithm follows the general and simple filter scheme, that is, it first evaluates the individual feature according to the evaluation criterion, and thereafter, the best n features are selected. However it uses a more complex evaluation function. The training samples, characterized by the selected features, are then passed to ID3. RELIEF collects discriminant information through local learning, and is solved as an eigenvalue decomposition problem with a closed-form solution. A fast implementation is also derived. Experiments on synthetic and real-world data are presented.

Mohanabharathi et al. [12] proposed an approach to perform the feature reduction and intrusion detection process under wireless LAN environment. The recurrent neural network is used for the intrusion detection process. The feature reduction process is also enhanced to improve accuracy. Real Time Recurrent Learning (RTRL) algorithm is used to solve the scalability problems and their approach is based on a hybrid approach, which combines the filter and wrapper models for selecting relevant features.

S. Mukherjee and N. Sharma [15] investigated the performance of three standard filter based feature selection methods using Correlation-based Feature Selection, Information Gain and Gain Ratio. The proposed method was Feature Vitality Based Reduction Method (FVBRM), to identify important reduced input features which reduces complexity and computation overheads. WEKA 3.6 a machine learning tool is used, to compute the feature selection subsets for CFS, IG, GR and FVBRM, and to measure the classification performance on each of these feature sets and choose the Naive Bayes classifier with full training set and10-fold cross validation for the testing purposes. The data set to be used for experiments is NSL-KDD labelled dataset. NSL-KDD dataset contains one type of normal data and 22 different types of attacks which falls into one of four categories. These are DoS, probe, R2L and U2R. They extracted only 62,986 records out of 1, 25,973 NSL-KDD dataset connections for training and testing. Empirical results showed that IG performs better than GR and FVBRM method shows much more improvement on classification accuracy as compared to CFS but takes more time.

K. Garg and P. Bhoria et al. [7] proposed a feature selection model for IDS which was applicable to detect only DOS attacks as DOS attacks affect the network badly so they must be detected on time but large set of 41 features for IDS reduces detection and classification efficiency so they irrelevant features must be reduced to increase efficiency of IDS for detecting DOS attacks on time. For this purpose they used NSL KDD dataset and analysis was performed on V26.1 data mining tool. Various selection methods were used for feature selection from 41 feature set and they were compared using C4.5 Decision Tree Classification algorithm.

V. Takellapati and G. Prasad [16] proposed an hybrid approach of feature selection for IDS in which IG and Triangle area based KNN were used for selecting features by combining Greedy K-mean clustering algorithm and SVM classifier to detect network attacks. This system achieved high accuracy detection rate and less error rate using KDD dataset for Intrusion Detection Systems. Polynomial based SVM to avoid problems caused by its high-dimensional feature space in classifying result was used here. But it needed to be implemented in real time web analytics for Intrusion detection.

Ahmed A. Elngar et al. [1] proposed an effective PSO-Discritize -HNB Intrusion Detection System in which Particle Swarm Optimization and Information Entropy Minimization discretize method with the Hidden Naïve Bayes Classifier and experiments were performed on KDD Dataset for Intrusion Detection Systems. Comparative study of applying IG with HNB Classifier was used to perform experiments and to validate the PSO-Discritize-HNB network intrusion detection system it was compared with various feature selection algorithms like Principal Component Analysis and Gain Ratio. The feature set was reduced to 11 which leads to high accuracy of 98.2% and improving speed to 0.18 seconds.

XI. CONCLUSIONS

In this paper, we have discussed all the basics of Data Mining process i.e. the requirements, the overall methodology and the concepts of Data Mining Process in brief. It’s very easy for a beginner to work upon data mining by going through the concepts that have been discussed in this paper. The Data mining Methodology and various techniques that are applicable in Data mining have been explained through this paper. Further, this paper could be considered to understand and explore the various kinds of research problems in Data Mining.

REFERENCES

[1] A. Elngar, A. Mohamed and F.M. Ghaleb. “A Real Time Anomaly Network Intrusion Detection System with High Accuracy.” *Information Sciences Letters, An International Journal,* May 2013.

[2] A. Sung, and S. Mukkamala. "The feature selection and intrusion detection problems." Advances in Computer Science-ASIAN 2004. Higher-Level Decision Making (2005): 3192-3193.

[3] A. Tamilarasan, S. Mukaamala, A.H. Sung and K.Yendrapalli."Feature Ranking and Selection for Intrusion Detection Using Artificial Neural Networks and Statistical Methods." International Joint Conference on Neural Networks: IEEE July 16-21 (2006).

[4] C. Lima, M. Assis and C. Protásio de Souza. "An Empirical Investigation of Attribute Selection Techniques based on Shannon, Rényi and Tsallis Entropies for Network Intrusion Detection." *American Journal of Intelligent Systems* 2.5 (2012): 111-117.

[5] Chebrolu, Srilatha and A. Abraham, and J. P. Thomas. "Feature deduction and ensemble design of intrusion detection systems." Computers & Security24.4 (2005): 295-307.

[6] Kdd cup 99 intrusion detection data set. Retrieved July 01, 2013.Online Available:http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

[7] K. Garg and P. Bhoria. “Determining Feature set of DOS attacks ”*International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE)* VOL. 3, Issue 5, May 2013.

[8] K. Khor, C. Ting and Somnuk. "A Feature Selection Approach for Network Intrusion Detection System.” *IEEE* (2009).

[9] L. Kuncheva "Combining Pattern Classifiers: Methods and Algorithms (Kuncheva, LI; 2004)[book review]." Neural Networks, IEEE Transactions on18.3 (2007): 964-964.

[10] M. Revathi and T. Rames. “Network Intrusion Detection System Using Reduced Dimensionality.” *Indian Journal of Computer Science and Engineering (IJCSE)* Vol.2, No .1(2010).

[11] Molina, L. Carlos, L. Belanche, and À. Nebot. "Feature selection algorithms: A survey and experimental evaluation." Data Mining, 2002. ICDM 2003. Proceedings. *2002 IEEE International Conference on*. IEEE, 2004

[12] Mohanabharathi, T.Kalaikumaran and S. Karthi “Feature Selection for Wireless Intrusion Detection System Using Filter and Wrapper Model.” *International Journal of Modern Engineering Research (IJMER)* Vol.2, Issue.4, July-Aug, 2012.

[13] P. Sangkatsanee, N. Wattanapongsakor and C .Charnsripinyo "Practical Real- time intrusion detection using Machine learning approaches.” Computer Communications (2011): 2227–2235.

[14] S Yang, K. Chi Chang , H. Wei and C. Lin. "Feature weighting and selection for a real-time network intrusion detection system based on GA with KNN." Intelligence and Security Informatics (2008): 195-204.

[15] S. Mukherjee and N. Sharma "Intrusion Detection using Naive Bayes Classifier with Feature Reduction" Procedia Technology 4 (2012) 119 – 128

[16] V.Takellapati and G. Prasad. "Network Intrusion Detection System based on Feature Selection and Triangle area Support Vector Machine." *International Journal of Engineering Trends and Technology (IJERT)* Vol.3, Issue 4, 2012.

[17] Witten. H .Ian and Frank .Eibe, P. 2005 Practical Machine Learning Tools and Techniques, 2nd edition.

[18] Data mining at http://www.zentut.com/data-mining "What id Data Mining."