Identification of reduced features for Intrusion Detection System using Naïve Bayes Classifier

Nisha R. Patil, Anand R. Wadhawane, Noopur V. Sarode, Sawan S. Rathod

Abstract— **In the era of new technologies there is a threat for security. Now-a-days there are various attacks or events occurring in the computer system. So, in order to detect this events occurring in the computer system we use Intrusion Detection. Intrusion detection mainly focuses on feature selection and the feature reduction. For identification of reduced features the system uses two techniques viz; Feature Selection Techniques and the Feature Vitality Based Reduction Method. This system uses Naive bayes Classifier on reduced datasets for identification of intrusions.**

Index Terms— **Intrusion Detection, NSL-KDD Dataset, FST, Naïve Bayes, Reduced Features.**

I. INTRODUCTION

In today's world the numbers of network based applications are developing rapidly in each and every sector like banking, military services, public web services etc. Thus, the use of internet has been increasing with the rapid development of network based applications. This increase has led to unauthorized activities. [1]These unauthorized activities are carried out by the external and internal attackers. The internal attackers are the fraud employees; they do this for the sake of their personal gain [2].

Intrusion is any kind of actions that consist of integrity, confidentiality and the availability of the resources. If the system fulfills this tokens i.e. integrity, confidentiality and the availability then the system is secured. Whereas, the intrusion detection is the process of observing and analyzing the events or the actions occurring in the computer system in order to detect the signs of security problems.

In this paper, the system performs the identification of reduced features for developing an Intrusion Detection System. For this, we make use of Feature Selection techniques like Information Gain and the Gain Ratio. This technique is used in identification of features. We also use the Feature Vitality Based Reduction Method (FVBRM method) for obtaining and identifying the reduced set of features which are important. The data mining algorithm named Naive Bayes Classifier is applied on the obtained reduced feature set for detection of intrusions. The result will show that the selected reduced attributes give better performance to design IDS that is efficient and effective for network intrusion detection.

II. RELATED WORK

It consist of study about Intrusion Detection System, Network Security terms, Data Mining, Feature Selection Techniques and Naive Bayes Classifiers.

The notion of intrusion detection was proposed by Anderson 1980's [2]. This described that audit trails contain valuable information and could be utilized for the purpose of misuse detection by identifying anomalous user behavior. Then the lead was taken by Denning at the SRI International and the first model of intrusion detection, 'Intrusion Detection Expert System' (IDES) was born in 1984 [3].A dynamic model "Intelligent Intrusion Detection System" proposed based on specific AI approach for intrusion detection. The techniques includes neural networks and fuzzy logic with network profiling, that uses simple data mining techniques to process the network data. The system combines anomaly, misuse and host based detection. Simple Fuzzy rules allow constructing if-then rules that reflect common ways of describing security attacks [4]. The accuracy and performance of IDS can be improved through obtaining good training parameters and selecting right feature to design any Artificial Neural Network (ANN) [5]. The feature ranking algorithm is used to reduce the feature space by using 3 ranking algorithm based on Support Vector Machine (SVM), Multivariate Adaptive Regression Splines (MARS) and linear Genetic programs (LPGs) [6].

In [9] author proposes "Enhanced Support Vector Decision Function "for feature selection, which is based on two important factors. First, the feature's rank, and second the correlation between the features. In [10], author proposes an automatic feature selection procedure based on Correlation –based Feature Selection (CFS). In [11] author investigate the performance of two feature selection algorithm involving Bayesian Network(BN) and Classification \& Regression Tee (CART) and ensemble of BN and CART and finally propose an hybrid architecture for combining different feature selection algorithms for intrusion detection. In [12], author proposes two phases approach in intrusion detection design. In the first phase, develop a correlation-based feature selection algorithm to remove the worthless information from the original high dimensional database. Next phase designs an intrusion detection method to solve the problems of uncertainty caused by limited and ambiguous information. In [13], Axellson wrote a well known paper that uses the Bayesian rule of conditional probability to point out that implication of the base-rate fallacy for intrusion detection. In [14], a behavior model is introduced that uses Bayesian techniques to obtain model parameters with maximal a-posteriori probabilities.

Manuscript received March 31, 2015.

Nisha R. Patil, Anand R. Wadhawane, Noopur V. Sarode, Sawan S. Rathod, Students of BE Computer, Sanjivani Collage of Engineering, Kopargaon, Savitribai Phule, Pune University .

III. SYSTEM OVERVIEW

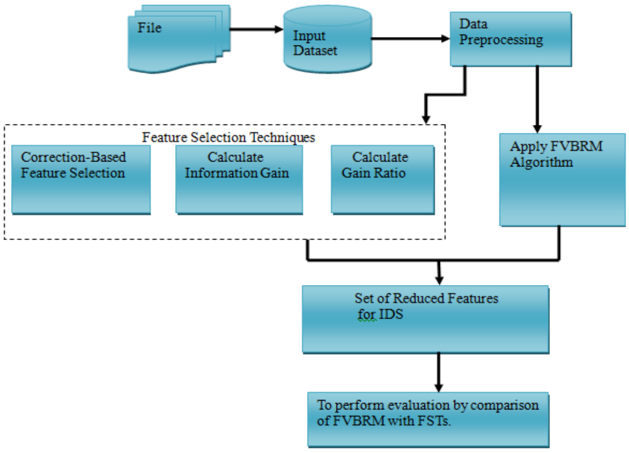

File
Input Dataset
Data Preprocessing
Feature Selection Techniques
Correction-Based Feature Selection
Calculate Information Gain
Calculate Gain Ratio
Apply FVBRM Algorithm
Set of Reduced Features for IDS
To perform evaluation by comparison of FVBRM with FSTs.

Fig:1 System Architecture.

IV. METHODOLOGY

We use three standard feature selection techniques for development of efficient and effective Intrusion Detection System. Those three techniques are Correlation-based Feature Selection (CFS), Information Gain (IG) and Gain Ratio (GR) to investigate important reduced input features. On the basis of discretizes values the reduced data sets are further selected by using common Naïve Bayes classifier. As results using discretizes features are generally more accurate, compact and shorter than using continuous values.

In this FVBRM method one input feature is deleted from the NSL-KDD 99 cup dataset at a time, the result we got is used for the training and testing of the classifier. This process continues until it performs better than the original dataset in terms of relevant, consistent and accurate performance criteria, known as Feature- Vitality Based Reduction Method (FVBRM).

A. INPUT DATASET

The data set used here is NSL-KDD labeled dataset. NSL-KDD dataset suggested solving some of the inherent problems of the KDD'99 data set. The numbers of records in the NSLKDD train and test sets are reasonable. This advantage makes it affordable to run the experiments on the complete set without the need to randomly select a small portion.

B. DATA PREPROCESSING

It is data mining technique that involves transforming raw data into an understandable format. Today's world databases are highly susceptible to noisy missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low quality data will lead to low quality mining results. There are several data preprocessing techniques as follows

a) Data Cleaning

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

b) Data Integration

Data mining often requires data integration—the merging of data from multiple data stores. Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can help to improve the accuracy and speed of the subsequent data mining process.

c) Data Reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

C. APPLY FEATURE SELECTION TECHNIQUES (FSTS)

It is an effective and an essential step in successful high dimensionality data mining application. The proposed system will use three features subset selection techniques like Correlation-based Feature Selection (CFS), Information Gain (IG), and Gain Ratio (GR). The overview of each one mentioned FST are given below

a) Information Gain (IG)

The IG evaluates attributes by measuring their information gain with respect to the class. It discretizes numeric attributes first using MDL based discretization method [7]. Let C be set consisting of c data samples with m distinct classes. [15]The training dataset ci contains sample of class I. Expected information needed to classify a given sample is calculated by

$$I(C_1, C_2 \; \dots \dots . . C_m) = - \sum_{i=1}^{m} \frac{C_i}{c} \log_2 \frac{C_i}{c}$$

Where $\frac{C_i}{c}$ is the probability that an arbitrary sample belongs to class Ci. Let feature *F* has v distinct values *{ f1, f2, …, fv}* which can divide the training set into v subsets *{C1,C2, …, Cv }* where Ci is the subset which has the value *fi* for feature *F*. Let Cj contain Cij samples of class *i*. The entropy of the feature *F* is given by

$$E(F) = \sum_{j=1}^{v} \frac{C_{ij} + \cdots + C_{mj}}{c} \times I(C_{ij}, \dots \dots . . C_{mj})$$

Information gain for F can be calculated as:

$$Gain(F) = I(C_1, \dots, C_m) - E(F)$$

EXAMPLE: Consider the following table.

Table 1. Class-Labeled Training Tuples from the Dataset

Duration	Protocol	Service	Flag	Src_bytes	Destn_bytes	Class
0	Tcp	ftp_data	SF	491	0	Anomaly
0	Udp	Other	SF	146	0	Anomaly
0	Tcp	Private	SO	0	0	Normal
0	Tcp	http	SF	232	8153	Anomaly
0	Tcp	http	SF	199	420	Anomaly
0	Tcp	Private	REJ	0	0	Normal
0	Tcp	Private	SO	0	0	Anomaly
0	Tcp	Private	SO	0	0	Normal
0	Tcp	Private	SO	0	0	Anomaly
0	Tcp	Private	SO	0	0	Anomaly
0	Tcp	Private	SO	0	0	Normal

Table1 represents a training set *D*, of class-labeled tuples randomly selected from the *Dataset* .In this example, each attribute is discrete valued. Continuous-valued attributes have been generalized. The class label attribute has two distinct values (namely Anomaly, Normal); therefore, there are two distinct classes (i.e., *m* =2). Let class *C*1 correspond to Anomaly and class *C*2 correspond to *Normal*. There are seven tuples of class *Anomaly* and four tuples of class *normal*. A (root) node *N* is created for the tuples in *D*. To find the splitting criterion for these tuples, we must compute
the information gain of each attribute.

Step1: First compute the expected information needed to classify a tuple in *D*

Info(D)=$-\sum_{i=1}^{m} \log_2 p_i$ --------(1)

By using this above formula we have to calculate *Info*(D).

Info(D)=$\frac{-7}{11}\log_2\left(\frac{7}{11}\right)-\frac{4}{11}\log_2\left(\frac{4}{11}\right)$

=$\frac{-7}{11}(-0.451)-\frac{4}{11}(-1.011)$

=0.287+0.3676

=0.6546 bits

Step2:

Next, compute the expected information requirement for each attribute. Let's start with the attribute **Protocol** .We need to look at the distribution of *Anomaly* and Normal tuples for each category of protocol. For the protocol category "Tcp," there are *six Anomaly* tuples and one Normal tuples. For the category "Udp," there is one Anomaly tuples and zero *Normal* tuples. Using Eq. (2),the expected information needed to classify a tuple in *D* if the tuples are partitioned according to *protocol* is

$Info_{Protocol}$ (D)=

$\frac{7}{11}\times\left(-\frac{6}{7}\log_2\left(\frac{6}{7}\right)-\frac{1}{7}\log_2\frac{1}{7}\right)+\frac{1}{11}\times\left(-\frac{1}{1}\log_2\frac{-1}{1}\right)$

=

$\frac{7}{11}\left(-\frac{6}{7}(-0.1541)-\frac{1}{7}(-1.9459)\right)+\frac{1}{11}\times(-1\times 0)$

=$\frac{7}{11}\left((0.1320)+(0.2279)\right)$

=$\frac{7}{11}(0.4099)$

=0.2609 bits.

Step 3:

Hence, the gain in information from such a partitioning would be

Gain(protocol)=Info(D)- $Info_{Protocol}$(D)

=0.6456-0.2609

=0.3937 bits.

Similarly we need to calculate all these factors for all the remaining attributes or sometimes only the number of the selected attributes.

b) Gain Ratio (GR)

The information gain measures prefer to select attributes having a large number of values. Gain ratio applies normalization to info gain using a value defined as

$SplitInfo_A(D)=-\sum_{j=1}^{v}\frac{|D_j|}{|D|}\times\log_2\left(\frac{|D_j|}{|D|}\right)$

------------------------Esq. (1)

The information gain measure is biased toward tests with many outcomes. That is, it Prefers to select attributes having a large number of values. For example, consider an Attribute that acts as a unique identifier such as *product_ID*. A split on *product_ID* would Result in a large number of partitions (as many as there are values), each one containing Just one tuple. Because each partition is pure, the information required to classify data set *D* based on this partitioning would be $Info_{product\_ID}$ *(D)* =0. Therefore, the information gained by partitioning on this attribute is maximal. Clearly, such a partitioning is useless for classification.

C4.5, a successor of ID3, uses an extension to information gain known as *gain ratio*, which attempts to overcome this bias.

The values in Esq.(1) represents the potential information generated by splitting the training data set, *D*, into *v* partitions, corresponding to the *v* outcomes of a test on attribute *A*. Note that, for each outcome, it considers the number of tuples having that outcome with respect to the total number of tuples in *D*. The gain ratio is defined as

Gain Ratio (A) =$\frac{\mathrm{Gain(A)}}{\mathrm{SplitInfo_A(D)}}$ ---------Eq (2)

The attribute with the maximum gain ratio is selected as the splitting attribute. Note, however, that as the split information approaches 0, the ratio becomes unstable. A constraint is added to avoid this, whereby the information gain of the test selected must be large—at least as great as the average gain over all tests examined.

Example: Computation of gain ratio for the attribute protocol

Refer the earlier **Table1.**

A test on *protocol* splits the data of Table into two partitions, namely Tcp and Udp containing ten and one tuples, respectively. To compute the gain ratio of protocol, we first use Esq. (1) to obtain

SplitInfo$_{protocol}$*(D)*

=$-\frac{10}{11}\times\log_2\left(\frac{10}{11}\right)-\frac{1}{11}\times\log_2\left(\frac{1}{11}\right)$

=-0.9090-(0.0954)-0.0909(-2.3979)

=0.0867+0.2179

=0.3046

From the example solved in information gain, we have *Gain*(protocol)=0.3937 bits. Therefore,

Gain Ratio(protocol)=0.3937/0.3046

=1.2925.

This is how we have to calculate gain ratio for each attribute (feature) from the selected dataset. In this example it is D.

D. *FVBRM ALGORITHM*

It is used to identify important reduced input features. The FVBRM algorithm works in following manner i.e. First, we will apply naïve bayes classifier on dataset with 41 features and its performance output like classifier's accuracy, RMSE, average TPR value and set F is input to this algorithm.

Let,
F=Full set of 41 features of NSL-KDD dataset
ac = classifiers accuracy
err = RMSE
avg_tpr= average TPR
// ac, err and avg_tpr resulted from invocation of NBC on full dataset, these values used as threshold values for feature selection

```
FVBRM Algorithm:
Begin
Initialize: S= {F}
For each feature {f} form
    1) T=S-{f}
    2) Invoke Naïve Bayes classifier on dataset with T features
    3) If CA>= ac And RMSE<=err And A_TPR>= avg_tpr then
       S=S-{f}
F=S // Set F with reduced features
End
```

NAÏVE BAYES CLASSIFIER:

Example:

Here, we wish to predict the class label of a tuple using naive Bayesian classification.

Refer the earlier **Table1.**

The data tuples are described by the attributes protocol, service, and flag. The class label attribute, Priority, has two distinct values (namely, Anomaly and Normal). Let *C*1 correspond to the class Anomaly and *C*2 Normal. The tuple we wish to classify is ***X***=(protocol=Tcp, service=private, flag=SF)

The prior probability of each class, can be computed based on the training tuples:

P(Priority=Anomaly)=7/11=0.63
P(Priority=Normal)=4/11=0.3636
P(X|C$_i$)
P(protocol=Tcp|Priority=Anomaly)=6/7=0.85
P(protocol=Tcp|Priority=Normal)=4/4=1
P (service=private |Priority=Anomaly)=2/7=0.28
P(service=private |Priority=Normal)=4/4=1
P(flag=SF |Priority=Anomaly)=4/7=0.57
P(flag=SF| Priority=Normal)=0/4=0

Using the above probabilities we get,

P(X|Priority=Anomaly)=P(protocol=Tcp|Priority=Anomaly) *P(service=private|Priority=Anomaly)*P(flag=SF|Priority= Anomaly)
=0.85*0.28*0.57
=0.1356

Similarly,

P(X|Priority=Normal)=P (protocol=Tcp |Priority=Normal)* P (service=private|Priority= Normal)* P (flag=SF|Priority= Normal)
=1*1*0
=0.

To find the class C$_i$ that maximizes *P (X|C$_i$***)** *P (C$_i$***)** we compute,

P(X|priority=Anomaly)P(priority=Anomaly)=0.1356*0.63= 0.0854. ----------- (1)

P(X|priority=Normal)P(priority=Normal)=0*0.3636
=0 ------------(2)

Here we can see that the tuple we considered is predicted that it belongs to Anomaly class(as the probability value for anomaly is greater than probability of normal from eqn 1 and 2.

V. RESULTS

Firstly perform the preprocessing.

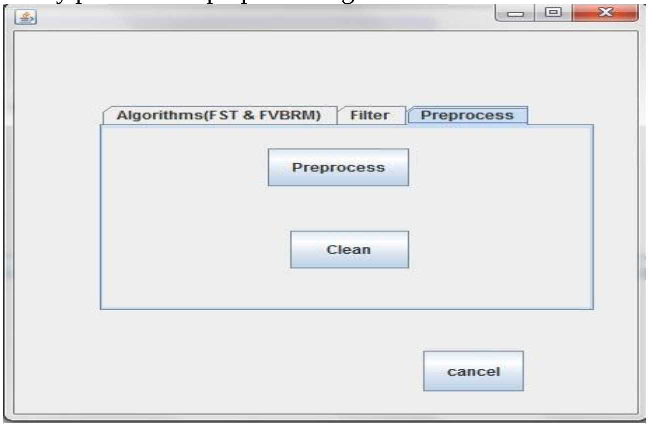
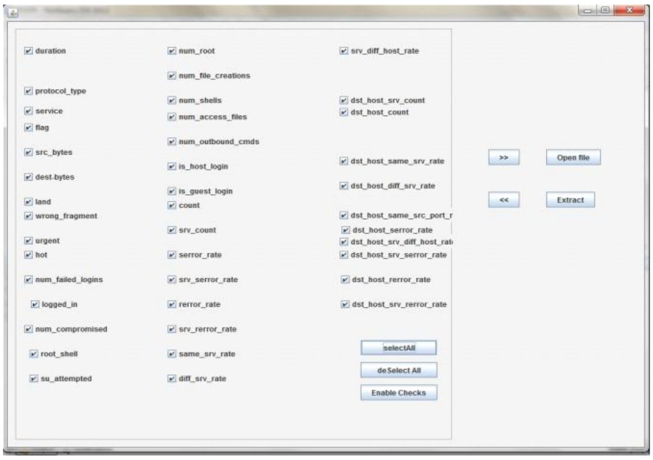


Figure 2: Homepage Snapshot

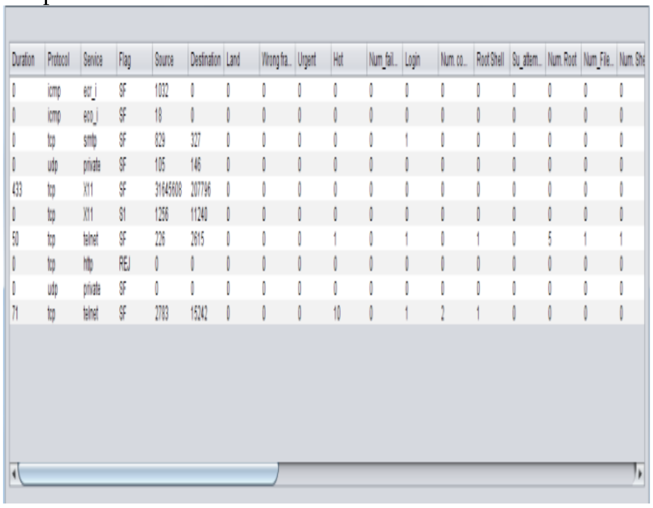


☑ duration
☑ protocol_type
☑ service
☑ flag
☑ src_bytes
☑ dest-bytes
☑ land
☑ wrong_fragment
☑ urgent
☑ hot
☑ num_failed_logins
☑ logged_in
☑ num_compromised
☑ root_shell
☑ su_attempted
☑ num_root
☑ num_file_creations
☑ num_shells
☑ num_access_files
☑ num_outbound_cmds
☑ is_host_login
☑ is_guest_login
☑ count
☑ srv_count
☑ serror_rate
☑ srv_serror_rate
☑ rerror_rate
☑ srv_rerror_rate
☑ same_srv_rate
☑ diff_srv_rate
☑ srv_diff_host_rate
☑ dst_host_srv_count
☑ dst_host_count
☑ dst_host_same_srv_rate
☑ dst_host_diff_srv_rate
☑ dst_host_same_src_port_r
☑ dst_host_serror_rate
☑ dst_host_srv_diff_host_rat
☑ dst_host_srv_serror_rate
☑ dst_host_rerror_rate
☑ dst_host_srv_rerror_rate

Figure 3: Feature Selections.

After selection of attributes now you need to open an input file by double click on "Open file" button. Select an input file from the "input files". On the selected input file perform the preprocessing.

Snapshot of Set of reduced features:



Duration	Protocol	Service	Flag	Source	Destination	Land	Wrong fra..	Urgent	Hot	Num_fail..	Login	Num co..	Root Shell	Su_attem..	Num Root	Num_File..	Num She
0	icmp	ecr_i	SF	1032	0	0	0	0	0	0	0	0	0	0	0	0	0
0	icmp	eco_i	SF	18	0	0	0	0	0	0	0	0	0	0	0	0	0
0	tcp	smtp	SF	829	327	0	0	0	0	0	1	0	0	0	0	0	0
0	udp	private	SF	105	146	0	0	0	0	0	0	0	0	0	0	0	0
433	tcp	X11	SF	31645608	207796	0	0	0	0	0	0	0	0	0	0	0	0
0	tcp	X11	S1	1256	11240	0	0	0	0	0	0	0	0	0	0	0	0
50	tcp	telnet	SF	226	2615	0	0	0	1	0	1	0	1	0	5	1	1
0	tcp	http	REJ	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	udp	private	SF	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	tcp	telnet	SF	2783	15242	0	0	0	10	0	1	2	1	0	0	0	0

VI. FUTURE SCOPE

There are various approaches being utilized in intrusion detections, but unfortunately any of the existing IDS so far is not completely flawless. So, the quest of betterment continues in the field IDS. The proposed system contains summarization study and identification of the drawbacks of formerly surveyed works. It is performed by using Feature Vitality Based Reduction Method which will be used to identify important reduced input features and one of the efficient classifier naive bayes on reduced datasets for intrusion detection.

VII. CONCLUSION

In this system, we are taking input dataset and preprocessing it by using various techniques like Data Cleaning, Data Integration, and Data Reduction. In Feature Vitality Based Reduction Method Naive Bayes Classifiers is applied on preprocessed data for feature selection. Feature selection Techniques like Information Gain and Gain Ratio is applied on preprocessed data. The output of both the methods i.e. Feature Vitality Based Reduction Method and three Feature selection Techniques like Information Gain and Gain Ratio is compared toget set of reduced features.

REFERENCES

[1] Dr.Saurabhi Mukherji, Neelam Sharma “Intrusion Detection using Naïve Bayes Classifier with Feature Reduction”.Department of computer science, Banasthali University, Jaipur, Rajasthan, 304022,India,2012

[2] James P. Anderson. Computer Security Threat Monitoring and Surveillance, 1980. http://csrc.nist.gov/publications/history/ande80.pdf.

[3] Dorothy E. Denning. An Intrusion Detection Model. IEEE Transactions on Software Engineering, 13(2):222–232, 1987.

[4] Norbik Bashah, Idris Bharanidharan Shanmugam, and Abdul Manan Ahmed,” Hybrid Intelligent Intrusion Detection System” World Academy of Science, Engineering and Technology, 2005.

[5] Saman M. Abdulla, Najla B. Al-Dabagh, Omar Zakaria, Identify Features and Parameters to Devise an Accurate Intrusion Detection System Using Artificial Neural Network, World Academy of Science, Engineering and Technology 2010.

[6] A. H. Sung, S. Mukkamala. The Feature Selection and Intrusion Detection Problems. In Proceedings of the 9th Asian Computing Science Conference, Lecture Notes in Computer Science 3029 Springer 2004.

[7] Mark A. Hall, Correlation-based Feature Selection for Machine Learning, Dept of Computer Science, University of Waikato. http://www.cs.waikato.ac.nz/~mhall/thesis.pdf.

[8] Wafa' S.Al-Sharafat, and Reyadh Naoum “Development of Genetic-based Machine Learning for Network Intrusion Detection” World Academy of Science, Engineering and Technology 55, 2009.

[9] S Zaman, F Karray Features selection for intrusion detection systems based on support vector machinesCCNC'09 Proceedings of the 6th IEEE Conference on Consumer Communications and Networking Conference 2009.

[10] H Nguyen, K Franke, S Petrovic Improving Effectiveness of Intrusion Detection by Correlation Feature Selection, International Conference on Availability, Reliability and Security, IEEE ,2010.

[11] S Chebrolu, A Abraham, J P. Thomas Feature deduction and ensemble design of intrusion detection systems, Computers & Security, Volume 24, Issue 4, June 2005.

[12] T. S. Chou, K. K. Yen, and J. Luo “Network Intrusion Detection Design Using Feature Selection of Soft Computing Paradigms. International Journal of Computational Intelligence, 2008.

[13] S. Axelsson, "The base rate fallacy and its implications for the difficulty of Intrusion detection", Proc. Of 6th. ACM conference on computer and communication security 1999.

[14] R. Puttini, Z.marrakchi, and L.Me, "Bayesian classification model for Real time intrusion detection”, Proc. of 22nd. International workshop on Bayesian inference and maximum entropy methods in science and engineering, 2002.

[15] Jiawei Han “Data Mining: Concepts and Techniques” Second Edition *University of Illinois at Urbana-Champaign* Michelin Kamber, 2006