Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 4 (2012) 119 – 128

Procedia Technology

C3IT-2012

Intrusion Detection using Naive Bayes Classifier with Feature Reduction

Dr. Saurabh Mukherjee[a], Neelam Sharma[a]

[a]*Department of Computer Science, Banasthali University, Jaipur,Rajasthan, 304022,India*

Abstract

Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems. Today most of the intrusion detection approaches focused on the issues of feature selection or reduction, since some of the features are irrelevant and redundant which results lengthy detection process and degrades the performance of an intrusion detection system (IDS). The purpose of this study is to identify important reduced input features in building IDS that is computationally efficient and effective. For this we investigate the performance of three standard feature selection methods using Correlation-based Feature Selection, Information Gain and Gain Ratio. In this paper we propose method Feature Vitality Based Reduction Method, to identify important reduced input features. We apply one of the efficient classifier naive bayes on reduced datasets for intrusion detection. Empirical results show that selected reduced attributes give better performance to design IDS that is efficient and effective for network intrusion detection.

Keywords: Network Security , Intrusion detection system, Feature selection , Data Mining, Naive Bayes classifier

1. Introduction

In recent years, due to dramatic growth in networked computer resources, a variety of network-based applications have been developed to provide services in many different areas, e.g., ecommerce services, public web services, and military services. The rapid growth of the Internet uses all over the world, estimation from the statistics up to December, 2010 [1], found that there are 6,845 Million Internet users worldwide. The increase in the number of networked machines has lead to an increase in unauthorized activity, not only from external attackers, but also from internal attackers, such as disgruntled employees and people abusing their privileges for personal gain[2].

An intrusion is defined as any set of actions that compromise the integrity, confidentiality or availability of a resource [3, 4]. If a system is able to assure that these three security tokens are fulfilled, it is considered secure. One of the biggest challenges in network-based intrusion detection is the extensive amount of data collected from the network. Thus the existing approaches of intrusion detection have focused on the issues of feature selection or dimensionality reduction. Feature selection or reduction keeps the original features as such and select subset of features that predicts the target class variable with maximum classification accuracy [5].

2212-0173 © 2012 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.protcy.2012.05.017

In this paper, the data mining algorithm naive bayes classifier will be evaluated on the NSL KDD dataset to detect attacks on the four attack categories: Probe (information gathering), DoS (deny of service), U2R (user to root) and R2L (remote to local). The feature reduction is applied using three standard feature selection methods Correlation-based Feature Selection (CFS), Information Gain (IG), Gain Ratio (GR) and our proposed model Feature Vitality Based Reduction Method (FVBRM). The naive bayes classifier's results are computed for comparison of feature reduction methods to show that our proposed model is more efficient for network intrusion detection. Rest of the paper is organized as follows: Section 2 and 3 give overview of IDS and Feature selection methods respectively. Section 4 describes naive bayes classifier. Related work is discussed in section 5. Section 6 provides s the research methodology of the work. The experimental setup discussed in section 8. The section 9 presents the result. Finally the paper is concluded with their future work in section 10.

2. Intrusion Detection System (IDS)

Intrusion is a type of attack that attempts to bypass the security mechanism of a computer system. Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems [6]. There are two main strategies of IDS [7]: misuse detection and anomaly detection. Misuse detection attempts to match patterns and signatures of already known attacks in the network traffic. A constantly updated database is usually used to store the signatures of known attacks. It cannot detect new attack until trained for them. Anomaly detection attempts to identify behavior that does not conform to normal behavior. This technique is based on the detection of traffic anomalies. The anomaly detection systems are adaptive in nature, they can deal with new attack but they cannot identify the specific type of attack.

Many researchers have proposed and implemented various models for IDS but they often generate too many false alerts due to their simplistic analysis. An attack generally falls into one of four categories:

- Denial-of-Service (DoS): Attackers tries to prevent legitimate users from using a service, these are smurf, neptune, back, teardrop, pod and land.
- Probe: Attackers tries to gain information about the target host. Port Scans or sweeping of a given IP-address range typically fall in this category (e.g. saint, ipsweep, portsweep and nmap).
- User-to-Root(U2R):Attackers has local access to the victim machine and tries to gain super user privileges, these are buffer_overflow, rootkit, landmodule and perl.
- Remote-to-Local(R2L): Attackers does not have an account on the victim machine, hence tries to gain access, these are guess_passwd, ftp_write, multihop, phf, spy, imap, warezclient and warezmaster.

3. Feature Selection

Feature selection is an effective and an essential step in successful high dimensionality data mining applications[8]. It is often an essential data processing step prior to applying a learning algorithm. Reduction of the attribute space leads to a better understandable model and simplifies the usage of different visualization technique. There are two common approaches for feature reduction. A Wrapper uses the intended learning algorithm itself to evaluate the usefulness of features, while filter evaluates features according to heuristics based on general characteristics of the data. The wrapper approach is generally considered to produce better feature subsets but runs much more slowly than a filter.

In this paper we are using three feature subset selection techniques Correlation-based Feature Selection(CFS), Information Gain (IG) and Gain Ratio(GR). Here, we are describing these techniques briefly.

3.1 Correlation-based Feature Selection (CFS)

CFS evaluates and ranks feature subsets rather than individual features. It prefers the set of attributes that are highly correlated with the class but with low intercorrelation[9]. With CFS various heuristic searching strategies such as hill climbing and bestfirst are often applied to search the feature subsets space in reasonable time. CFS first calculates a matrix of feature-class and feature-feature correlations from the training data and then searches the feature subset space using a bestfirst. Equation 1 (Ghiselli 1964) for CFS is

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

Where M_s is the correlation between the summed feature subset S, k is the number of subset feature, $\overline{r_{cf}}$ is the average of the correlation between the subsets feature and the class variable, and $\overline{r_{ff}}$ is the average inter-correlation between subset features[10].

3.2 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[9] . Let C be set consisting of c data samples with m distinct classes. The training dataset ci contains sample of class I. Expected information needed to classify a given sample is
calculated by:

$$I(c_1, c_2, \ldots\ldots\ c_m) = -\sum_{i=1}^{m} \frac{c_i}{c} \log_2\left(\frac{c_i}{c}\right)$$

Where $\frac{c_i}{c}$ is the probability that an arbitrary sample belongs to class C_i. Let feature *F* has v distinct values $\{f_1, f_2, \ldots, f_v\}$ *which* can divide the training set into v subsets $\{C_1, C_2, \ldots, C_v\}$ where C_i is the subset which has the value f_i for feature *F*. Let C_j contain C_{ij} samples of class *i*. The entropy of the feature *F* is given by

$$E(F) = \sum_{j=1}^{v} \frac{C_{1j} + \ldots + C_{mj}}{c} \times I(c_{ij}\ \ldots\ldots\ldots c_{mj})$$

Information gain for F can be calculated as:

$$Gain(F) = I(c_1,\ldots,c_m) - E(F)$$

3.3 Gain Ratio (GR)

The information gain measures prefers to select attributes having a large number of values. The gain ratio an extension of info gain, attempts to overcome this bias. Gain ratio applies normalization to info gain using a value defined as

$$SplitInfo_F(C) = -\sum_{i=1}^{v} (|C_i|/|C|) \log_2(|C_i|/|C|)$$

The above value represents the information generated splitting the training data set C into v partitions corresponding to v outcomes of a test on the feature F [11].
The gain ratio is defined as

$$Gain\ Ratio(F) = Gain(F) / SplitInfo_F(S)$$

4. Naïve Bayes Classifier

The naïve Bayes model is a heavily simplified Bayesian probability model[12]. The naïve Bayes classifier operates on a strong independence assumption [12]. This means that the probability of one attribute does not affect the probability of the other. Given a series of n attributes,the naïve Bayes classifier makes 2n! independent assumptions. Nevertheless, the results of the naïve Bayes classifier are often correct. The work reported in[13] examines the circumstances under which the naïve bayes classifier

performs well and why. It states that the error is a result of three factors: training data noise, bias, and variance. Training data noise can only be minimised by choosing good training data. The training data must be divided into various groups by the machine learning algorithm. Bias is the error due to groupings in the training data being very large. Variance is the error due to those groupings being too small.

5. Related Work

The notion of intrusion detection was born in the 1980's with a paper from Anderson[14]. which described that audit trails contain valuable information and could be utilized for the purpose of misuse detection by identifying anomalous user behaviour. The lead was then taken by Denning at the SRI International and the first model of intrusion detection, 'Intrusion Detection Expert System' (IDES) [15] was born in 1984.

In [16], 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. There have been many techniques used for machine learning applications to tackle the problem of feature selection for intrusion detection. In [17], author used PCA to project features space to principal feature space and select features corresponding to the highest eigen values using Genetic Algorithm. In [18] author shows that the accuracy and performance of an IDS can be improved through obtaining good training parameters and selecting right feature to design any Artificial Neural Network (ANN). In [19] author used feature ranking algorithm to reduce the feature space by using 3 ranking algorithm based on Support Vector Machine (SVM) , Multivariate Adaptive Regression Splines (MARS) and linear Genetic programes (LPGs). In [20] author propose "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 [21], author propose an automatic feature selection procedure based on Correlation –based Feature Selection (CFS). In [22] 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 [23], 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 [24], 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 [25], a behaviour model is introduced that uses Bayesian techniques to obtain model parameters with maximal a-posteriori probabilities.

6. Research Methodology

For building efficient and effective IDS we investigate the performance of three standard feature selection algorithms involving Correlation-based Feature Selection (CFS), Information Gain (IG) and Gain Ratio (GR) to identify important reduced input features. The reduced data sets are further classified by using common Naïve Bayes classifier on discretized values. Since results using discretized features are usually more compact, shorter and accurate than using continuous values.

In proposed FVBRM, one input feature is deleted from the dataset at a time , the resultant dataset is then used for the training and testing of the classifier, this process continues until it performs better than the original dataset in terms of relevant performance criteria, known as Feature- Vitality Based Reduction Method(FVBRM) in this paper. The metrics will show that the FVBRM performs much better than other feature reduction methods like CFS, IG and GR. The feature reduction is performed on 41 features to get 10 using CFS, 14 using GR, 20 using IG and 24 using FVBRM on NSL-KDD dataset. The empirical results are compared for different feature reduction methods using identified performance metrics like

classification Accuracy, Root Mean Squared Error(RMSE), True Positive Rate(TPR) for attack class values (dos, probe. R2l, u2r and normal), time taken to build model and confusion matrix.

7. Proposed Method

In this approach we have achieved the subset of 24 features by reducing NSL-KDD [26] dataset of 41 features for intrusion detection on the basis of feature's vitality. The vitality of feature is determined by considering three main performance criteria the classification accuracy, TPR and FPR of the system. We used sequential search to identify the important set of features: starting with the set of all features, one feature was removed at a time until the accuracy of the classifier was below a certain threshold. In other words, the feature selection of is "leave-one-out" remove one feature from the original dataset, redo the experiment, then compare the new results with the original results. Since there are 41 features in NSL-KDD data set, the experiment is repeated 41 times to ensure that each feature is either important, unimportant or less important. By deletion of a feature if performance decreases then feature is important, if performance increases then feature is unimportant and if no changes found in performance then feature is less-important. Here we have explained the algorithm for FVBRM. First, we 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.

Input:
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

8. Experimental Setup

We used WEKA 3.6 a machine learning tool [27], to compute the feature selection subsets for CFS, IG ,GR and FVBRM, and to measure the classification performance on each of these feature sets. We choose the Naïve Bayes classifier with full training set and 10-fold cross validation for the testing purposes. In 10-fold cross-validation, the available data is randomly divided into 10 disjoint subsets of approximately equal size. One of the subsets is then used as the test set and the remaining 9 sets are used for building the classifier. The test set is then used to estimate the accuracy. This is done repeatedly 10 times so that each subset is used as a test subset once. The accuracy estimates is then the mean of the estimates for each of the classifiers. Cross-validation has been tested extensively and has been found to generally work well when sufficient data is available.

Dataset Description

The data set to be used in our experiments is NSL-KDD labeled dataset. NSL-KDD dataset suggested to solve some of the inherent problems of the KDD'99 data set[25].The number of records in the NSL-KDD 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. Consequently, evaluation results of different research work will be consistent and comparable. 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. we extracted only 62,986 records out of 1,25,973 NSL-KDD dataset connections for training and testing.

Table 1. Exemplify distribution of classes and the percentage of attacks

Category of Attack Class (Class)	Number of instances/records	Percentage of Class Occurrences (Approximate)
Normal	33896	54%
DoS	22817	36%
Probe	5781	9%
U2R	25	0.03%
R2L	467	0.7%
Total	62,986	100%

Table 1 shows the distribution of classes in the actual training data for classifiers evaluation and the percentage of attacks is displayed using bar chart in Fig 1.

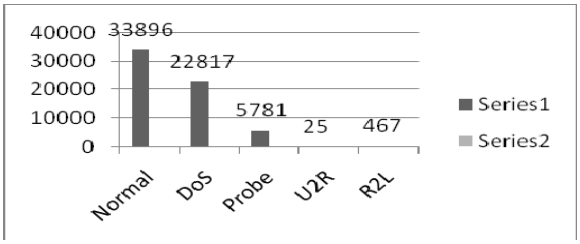

Figure 1

9. Results

We used three standards and one proposed model for feature reduction. The feature reduction performed on 41 features and obtained 10, 14 20 and 24 by CFS, GR, IG and FVBRM respectively.

Table2. Depicts the number of features selected by each feature reduction method.

Feature Selection Technique	# Attribute Selected	Selected Attributes
CFS+ BestFirst	10	3,4,5,6,12,26,29,30, 37,38
GR + Ranker	14	3,4,5,6,11,12,22,25,26,29,30, 37, 38,39
InfoGain + Ranker	20	3,4,5,6,12,23, 24 ,25,26,29,30, 31 ,32,33,34,35 , 36 ,37,38,39
FVBRM	24	1,3,4,5,6,7,8,9, 10,11, 12,13,14, 15,16,17, 18, 19 ,23,24,32,33, 36,38,40

Performance Evaluation

To evaluate the results of classifier, we have used standard metrics such as confusion matrix, true positive rate, false positive rate, and classifier's accuracy.

Confusion Matrix- This may be used to summarize the predictive performance of a classifier on test data. It is commonly encountered in a two-class format, but can be generated for any number of classes. A single prediction by a classifier can have four outcomes which are displayed in the following confusion matrix.

Table3. Show confusion matrix.

Actual Class	Predicted Class	
	Class=Yes	Class=No
Class=Yes	TN	FP
Class=No	FP	TN

True Positive (TP), the actual class of the test instance is positive and the classifier correctly predicts the class as positive. False Negative (FN), the actual class of the test instance is positive but the classifier incorrectly predicts the class as negative. False Positive (FP), the actual class of the test instance is negative but the classifier incorrectly predicts the class as positive. True Negative (TN), the actual class of the test instance is negative and the classifier correctly predicts the class as negative.

True Positive Rate(TPR) or *Sensitivity* or *Recall (R)* is defined as: $TPR=TP/(TP+FN)$

False Positive Rate(FPR) is: $FPR=FP/(TN+FP)$

We can obtain the accuracy of a classifier by $Accuracy =(TP+TN)/(TP+FN+FP+TN) * 100\ \%$

Figure 2 shows comparative graph for classification accuracy (%) for Naïve Bayes classifier on reduced feature's dataset obtained by (i) CFS+BestFirst (ii) GR+Ranker (ii) IG+Ranker (iv) FVBRM and on dataset with all features.

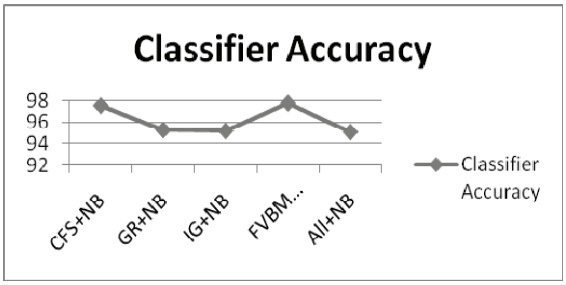

Figure 2

Figure 3 shows detection rate for dos, probe, r2l ,u2r and normal network connections for various reduced datasets achieved by different feature selction techniques

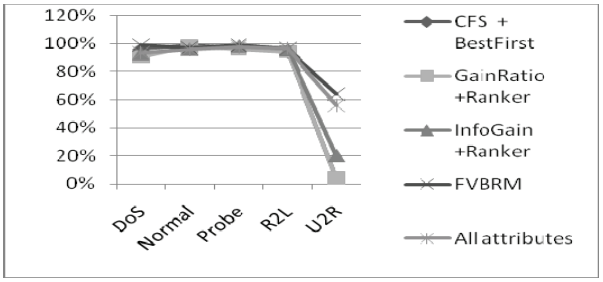

Figure 3

Table4. Depicts the false positive rate for dos, probe, r2l and u2r.

Feature Reduction Methods→ / Attack Types ↓	CfsSubset AttributeEval + BestFirst	GainRatio AttributeEval +Ranker	InfoGain AttributeEval+Ranker	**Feature Vitality Based Method**	All attributes
DoS	0.002	0.005	0.001	**0.001**	0.001
Probe	0.013	0.034	0.02	**0.011**	0.024
R2L	0.001	0.002	0.01	**0.004**	0.008
U2R	0.002	0.002	0.004	**0.005**	0.007

The empirical results in table 5(given on page 9) clearly indicate that no existing feature reduction method perform best for intrusion detection. Although the reduced feature set obtained in FVBRM is the largest one but it performs better than other methods. The FVBRM method achieved 97.78% overall classifier's accuracy with 98.7 TPR for DoS, 97%for normal, 98.8% for probe, 96.1 for r2l and 64% for u2r which is the highest as compared to others.

Table 5. Indicate feature reduction methods performance.

Statistical Result→					**TPR for Attack Class values**				
Feature Reduction Methods↓	**Attribute #**	**Taken to Build Model**	**Accuracy**	**Root Mean Squared Error**	**DoS**	**Normal**	**Probe**	**R2L**	**U2R**
CfsSubsetAttributeEval + BestFirst	10	6.81	97.55	0.0885	96%	98.6%	97%	94%	4%
GainRatioAttributeEval +Ranker	14	4.3	95.30	0.1138	90.6%	98.4%	96%	94.2%	4%
InfoGainAttributeEval +Ranker	20	8.13	95.21	0.1262	93%	96.2%	98.6%	95.9%	20%
FVBRM	**24**	**9.42**	**97.78**	**0.083**	**98.7%**	**97%**	**98.8%**	**96.1%**	**64%**
All attributes	41	16.41	95.11	0.1274	93.5%	95.7%	97.8%	96.8%	56%

10. Conclusion and Future Work

In this paper we have proposed FVBRM model for feature selection and make its comparison with three feature selectors CFS, IG and GR. Experimental result illustrates feature subset identified by CFS

has improved Naïve Bayes classification accuracy when compared to IG and GR. Although GR is an extended of IG , but in our analysis we have used both the techniques for feature selection and IG performs better than GR. FVBRM method shows much more improvement on classification accuracy with compared to CFS but takes more time. Future work will include customize of FVBRM feature selection method to improve the results for intrusion particularly for U2R attacks with reduced complexity and overheads.

Reference:

1. http://www.internetworldstats.com/stats.htm
2. Richard Power. 1999 CSI/FBI computer crime and security survey. Computer Security Journal, Volume XV (2), 1999.
3. Jian Pei Shambhu J. Upadhyaya Faisal Farooq Venugopal Govindaraju. Proceedings of the 20th International Conference on Data Engineering (ICDE'04) 1063-6382/04 $ 20.00 © 2004 IEEE
4. Debar, H., Dacier, M., and Wespi, A., A Revised taxonomy for intrusion detection systems, Annales des Telecommunications, Vol. 55, No. 7–8, 361–378, 2000.
5. R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence, 1(2) (1997) 273–324.
6. Bace, R. (2000). Intrusion Detection. Macmillan Technical Publishing.
7. Markou, M. and Singh, S., Novelty Detection: A review, Part 1: Statistical Approaches, Signal Processing, 8(12), 2003, pp. 2481-2497.
8. Liu H ,Setiono R, Motoda H, Zhao Z Feature Selection: An Ever Evolving Frontier in Data Mining, JMLR: Workshop and Conference Proceedings 10: 4-13 The Fourth Workshop on Feature Selection in Data Mining.
9. I.H.Witten, E.Frank, M.A. Hall " Data Mining Practical Machine Leanrning Tools & Techniques" Third edition, Pub. – Morgan kouffman.
10. 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.
11. j.Han ,M Kamber, Data mining : Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers(2001).
12. 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
13. Ms.Nivedita Naidu, Dr.R.V.Dharaskar "An effective approach to network intrusion detection system using genetic algorithm", International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 2, 2010.
14. James P. Anderson. Computer Security Threat Monitoring and Surveillance, 1980. Lastaccessed: Novmeber 30,2008 . http://csrc.nist.gov/publications/history/ande80.pdf.
15. Dorothy E. Denning. An Intrusion-Detection Model. IEEE Transactions on Software Engineering,13(2):222–232, 1987. IEEE.
16. Norbik Bashah, Idris Bharanidharan Shanmugam, and Abdul Manan Ahmed," Hybrid Intelligent Intrusion Detection System" World Academy of Science, Engineering and Technology, 2005
17. I Ahmad, A B Abdulah, A S Alghamdi, K Alnfajan,M Hussain, Feature Subset Selection for Network Intrusion Detection Mechanism Using Genetic Eigen Vectors, Proc .of CSIT vol.5 (2011)
18. 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.
19. A. H. Sung, S. Mukkamala. (2004) The Feature Selection and Intrusion Detection Problems. In Proceedings of the 9th Asian Computing Science Conference, Lecture Notes in Computer Science 3029 Springer 2004, pp.

20. 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
21. H Nguyen, K Franke, S Petrovic Improving Effectiveness of Intrusion Detection by Correlation Feature Selection, 2010 International Conference on Availability, Reliability and Security,IEEE Pages-17-24
22. S Chebrolu, A Abraham, J P. Thomas Feature deduction and ensemble design of intrusion detection systems, Computers & Security,Volume 24, Issue 4, June 2005, Pages 295-307
23. 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 4;3 2008
24. 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.
25. 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.
26. NSL-KDD dataset for network –based intrusion detection systems" available on http://iscx.info/NSL-KDD/
27. http://www.cs.waikato.ac.nz/~ml/weka/