A Feature Covariance Deviation Method for Feature Reduction in Intrusion Detection

B.V. RamNaresh Yadav, B.Satya Narayana, and D. Vasumati

Abstract**—** Data security is primary concern in all service providing systems. Intrusion detection system is being popularly used for safeguard the data. But, traditional intrusion detection systems are based on derived knowledge of signature of known attacks which limit the scope of intrusion detection. The wide use of internet and its services in today life make high dependency over computer network and Web services systems. The dependency demands for a high network security for the exchange of confidential and secure information over the network communication channel. A secure information exchange can be made through deploying efficient intrusion detection for protection from various network attacks. This paper proposes a feature reduction approach based on feature covariance deviation method (FCDM) and a modified naïve Bayesian algorithm for efficient classification in intrusion detection. Evaluation measures of the proposed reduction method is performed in compare with other feature reduction methods and classification approaches shows a better performance using NSL-KDD data set.

Keywords**—**Feature Reduction, NSL-KDD, Classification, Intrusion Detection, FCVM, Naïve Bayesian.

I. Introduction

As the increasing needs of the internet in everyday life and our dependence over the web services systems over distributed computer networks demands network security as a necessary condition of the world to receive confidential information. Most of the sensitive information is high prone to the attacks as they are specially targeted by attackers. The cause of high intrusion may be due to internal and external vulnerabilities activities with a system. The vulnerabilities activities might be occur because of security breaches, improper configuration or program execution. It is also possible attacker can perform multiple vulnerability combination to intrude which create challenge for detection. To make a system secure from attacker's intrusion detection systems play a vital role in diversified network systems [1][16]. The vast and distributed network consist of high number of distributed services running in many online servers, these networks services are more open for attackers. To prevent high efficient intrusion detection system are needed in network system.

Based on the detection pattern intrusion detection are generally identified in two categories, Anomaly and Misuse. Misuse attack detection is based on the system knowledge over the past vulnerabilities patterns, where as anomaly attack detection is performed based on pattern deviation in compare to normal patterns. To detection these two categories of attacks many intelligent approaches are proposed and applied. Some of them are based on the pattern matching, transition analysis, rule-based identification and genetic approach for misuse attack detection[2][5] and statistical analysis, inductive sequential pattern analysis, artificial neural network and various other data mining approaches are used for anomaly attacks detection[23]. A major problem for IDS is that it can give false alarms [22] in cases a small modification in the normal system behavior. The IDS must be capable of adapting to these changes and the detection pattern must be updated in regular intervals. One straight forward approach can be used to generate a new pattern with each set of new audited data to incorporate patterns of intrusion behavior.

This paper proposes a feature covariance deviation method (FCDM) for feature reduction and also modified the naïve bayesian algorithm for efficient classification using NSL-KDD dataset. We describe the proposal in the following sections. The sections of the paper are organized as; Section-2 presents related works, Section-3 presents the proposed feature covariance deviation method for efficient feature selection and Section-4 presents the modified naïve bayesian algorithm for classification, Section-5 describe of NSL-KDD dataset and experiment result analysis in section-6 and section-7 presents the paper conclusion.

II. Related Works

Many researchers have proposed and implemented different intrusion detection models that define different measures of system performance with an ad hoc assumption that normality and abnormality manifested precisely on the selected features sets for modeling and analysis. In [12][13][14] some approaches are defined on building and analyzing of intrusion detection system.

Zubair A.Baig et al. [7] present computer network security model using AODE-based for Intrusion Detection System and the study and observation suggest that the Naive Bayes intrusion detection does not accurately detect network intrusions and required improvisation. Panda M. et al., [8] performs a series of experiment study and observes the accuracy and performance measures of Naïve Bayes Classifier

B.V.RamNaresh Yadav is with the JNTUHCE Nachupally, Karimnagar, Telangana State, India (+91-9490685386; e-mail: bvramnaresh@gmail.com).

B.Satya Narayana was with SK University, Ananthapur, Andhra Pradesh, India (e-mail: bachalasatya@yahoo.com).

D.Vasumathi is with the JNTUHCEH, Hyderabad, Telangana State, India (e-mail:rochan44@gmail.com).

http://dx.doi.org/10.15242/IIE.E0415046

in compares with the different mining approaches for intrusion detection system in all classes. It shows a better accuracy in compare to decision tree approach but it also concludes that decision tree approach is better in case of anomaly detection.

Ektefa M et al., [11] perform a comparison analysis study between C4.5 and SVM [17]. It evaluates the comparison study using KDD'99 datasets. The study conclude that network intrusions and false alarm detection is better in C4.5 in compare to SVM approach and Hai Nguyen et al., [9] also performed a comparison study between C4.5 and BayesNet using KDD'99 datasets. Wei Lu et.al [8] performs experimental study over NSL-KDD[2] datasets using various machine learning algorithms and achieve better accuracy result in compare to KDD'99[20] datasets as NSL-KDD datasets are filtered from redundant data and use separate datasets for training and testing which provide high accuracy in intrusion detection.

M Jianliang [10] performs intrusion detection analysis using unsupervised learning approaches and K-mean algorithm for datasets clustering. J Zhang and M Zulkernine [15] perform anomaly based intrusion detection using random forest algorithm. Gary Stein [18] performs feature reduction and intrusion detection with genetic algorithm and decision tree algorithm. Cuixio Z et al., [19] perform anomaly and misuse detection through designing a mixed approach using missed detection model build using unsupervised clustering methods.

Studies illustrate that most of the researchers had used KDD'99 [20] datasets for the practical evaluation, which suffers from drawback of redundant data. Most of the previous works had implements single method approach or cross validation data sets for detecting multiple attacks types based on the known attacks. With multiple machine learning approaches [2][5] using KDD'99 datasets many studies are made but no efforts are made to improve accuracy through feature selection measures. This motivates us to use NSL-KDD and propose more efficient method with effective feature selection and classification to achieve high accuracy and fewer false alarms in intrusion detection.

III. Proposed Feature Covariance Deviation Method

Feature reduction is a challenging issue in intrusion detection as high data redundancy is appears in datasets. Redundant data is another issue in data integration and correlation. A feature may have abundant redundant if it is generated from inconsistent resources. Network attackers generate such kinds of abundant redundant data consistently with a least feature variance which creates a high challenge in intrusion detection and cause of false alarms. A high number of feature comparisons for intrusion detection affect the intrusion detection system. We propose a feature reduction method based on feature covariance deviation (FCD) to provide relevant features which will be effective for intrusion detection.

Covariance is mostly used to measure the standard deviation between two or more features. It is useful in finding the change in one feature corresponding to the average amount of change in the other feature. Usually we want to find the possible relationship and deviation between two features in which the values of one feature are affected by the values of the other. The degree of relation deviation between two such sets of features is measured by covariance deviation as σ.

Given two features can be measure how strongly one implies the other based on the available data. In probability theory and statistics correlation and covariance are two similar measures for assessing how many two attributes changes together. Consider two numeric attributes *A* and *B*, and a set of *n* unique data observations $\{(a_1, b1), \ldots, (a_n ,b_n)\}$. The mean values of *A* and *B*, respectively, are also known as the expected values on *A* and *B*, that is,

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

and,

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}$$

the covariance deviation, σ_A between *A* with other features is defined as,

$$\sigma_A = \sum_{i=i+1}^{n} E(A) - E(B_i)$$

Based on the covariance deviation, σ between the features we create two sets of deducted features sets. If σ is less then 0, it means that the feature has less variation in data collection and have less impact on intrusion detection, and if σ is more than 0 or higher make the features high variations and create more difficult on intrusion detection. We collect a sets of features which have σ >=1 for our experiment evaluation.

IV. Modified Naïve Bayseian Alogirthm

Naïve Bayesian classifiers works on assumption that a class is free from it feature values variations, this assumption is generally called as condition independence. This approach is made for the computation simplification in relate to "Naïve". It makes classifiers to represents the dependencies of subsets of attributes in relate to their class. It was observed that bayesian approach is effective in certain situation and it's highly dependent on the assumptions of the target system information for the efficient results. Due to high dependency a small deviation in the assumption hypothesis makes a lot of errors in detection [3].

We modified the Naïve Bayesian using feature covariance deviation method to obtain reduce feature and efficiently classifying the datasets as proposed in Alogorithm-1 below.

```
Algorithm-1: Modified Naïve Bayes classification
Input:
    $S$ → Set of Training dataset
    $C$ → Set of attack category
    Cat_Data[ ] → Empty set

Method1: Reduce_Features()
    For each record data $r_i$ of $S$
```

```
        For each category data of $c_i$ of $C$
          If $c_i == C[r_i]$ then
         $Cat\_Data[\ ] \leftarrow c_i$
          End if
End For
  End For
 $Reduct\_Features[\ ] \rightarrow$ Empty set
 For each category data $c_i$ in $Cat\_Data[\ ]$
$Feature\_Data[\ ] \leftarrow c_i$
 For each feature data $f_i$ in $Feature\_Data[\ ]$
 Find covariance deviation, $\sigma$ for each feature in $f_i$ in
compare to $f_{i+1}$
   If $\sigma >= 1$ then
       $Reduct\_Features[\ ] \leftarrow f_i$
   End if
 End For
```

Method2: ***Data_Classification ()***

```
Input:
    $T \rightarrow$ Set of Test dataset
    $Trained\_FeatureSet[\ ][\ ] \rightarrow$ n-dimensional vector of
$Reduct\_Features[\ ]$
    $Reduct\_Features[\ ] \rightarrow$ Set of Reduced Features

  For each data record $t_i$ of $T$
  For each feature data of $f_i$ in $Reduct\_Features[\ ]$
  For of each $Trained\_FeatureSet[c_i\ ][\ ]$ of $Cat\_Data[\ ]$
   Compute the Bayes probability similarity, $\beta$ of $f_i$ in
$Trained\_FeatureSet[c_i\ ][\ ]$
          End for
 End for
       If $\beta >= 1$ then
   $t_i$, classified as $\rightarrow c_i$
       End if
   End for
```

Using the above proposed feature deduction and classification approach we perform a regressive testing over the set of test data of NSL-KDD for the evaluation in compare with other feature reduction method like Information Gain (IG) and Gain Ration (GR) with Bayes classification.

V. Description of NSL-KDD Dataset

This paper use NSL-KDD is dataset for the evaluation, it resolve inherent problems of data redundancy of the KDD'99 data set which are presented in [21]. The dataset consists of normal data and four category of network attacks as described in Table-1.

TABLE I
Attacks Category and Types

Category	Types
DOS	Apache2, Back, Land, Mailbomb, SYNFlood, Processtable, Smurf, Teardrop, Udpstrom
Probe	IPsweep, Mscan, Nmap, portsweep, Saint, Satan
R2L	Guesspasswd, Ftpwrite, Imap, multihop, Named, Phf, Sendmail, snmp, getattack, snmpguess, warezmaster, worm, Xlock, Xnsoop
U2R	Bufferoverflow, http, tunne, Loadmodule, perl, rootkit, ps, sqlattack, xterm

NSL-KDD data set filtered out the redundant data records in its training and testing datasets, which helps classifiers to perform better classification. The collection of sets being used for the proposed test does not contains any duplicate records which helps in the improvisation of the performance of training process and gets better detection rates in testing. We have performed our training data extraction on an approx of 20% of the NSL-KDD datasets which is only 37,040 records and for testing 22,544 records. The extracted datasets classes based on the category are shown in Table-2 and in Figure-1.

TABLE II
Training Data Category Class Distribution

CATEGORY	NO. OF RECORDS	% OF CLASS
Normal	15601	42.12%
DOS	13574	36.65%
PROBE	4692	12.67%
U2R	213	0.58%
R2L	2962	8.00%
TOTAL:	37042	

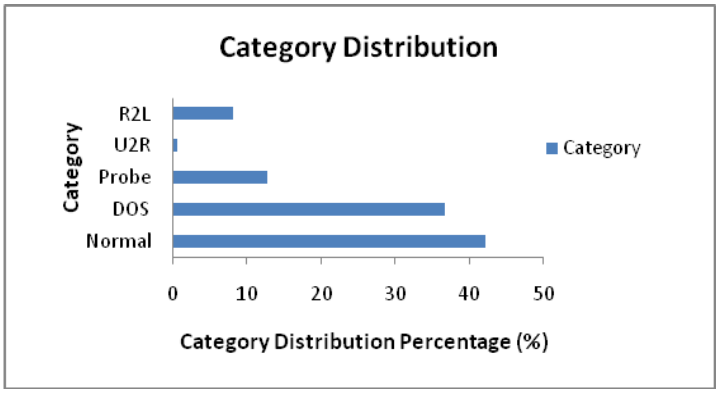

Fig.1 Training Data Category Class Distribution

VI. Experiment Result Analysis

To perform an experiment analysis of our proposal we use Java and WEKA 3.6 tool. Initially we implement feature selection using FCDM, Info Gain (IG) and Gain Ration (GR) method over trained datasets to obtain the reduced feature sets as shown in Table-3.

TABLE III
Reduced Features Sets

Feature Selection Methods	Features Selected	Features Count
FCDM	F-1, F-5, F-6, F-23, F-24, F-25, F-26, F-27, F-28, F-29, F-31, F-32, F-33, F-34, F-35, F-36, F-37, F-38, F-39, F-40, F-41	21
InfoGain +Ranker	F-3, F-4, F-5, F-6, F-12, F-23, F-24, F-25, F-26, F-29, F-30, F-31, F-32, F-33, F-34, F-35, F-36, F-37, F-38, F-39	20
GainRation +Ranker	F-3, F-4, F-5, F-6, F-11, F-12, F-22, F-25, F-26, F-29, F-30, F-37, F-38, F-39	14

The obtained features selected based on the feature selection methods we analyze the Classification Accuracy (CA), Root Mean Square Error (RMSE) and True positive rate (TPR) for different attack category to measure the effectiveness of the proposal.

http://dx.doi.org/10.15242/IIE.E0415046

A. True and False Positive Rate

To compute TPR and FPR we have used a standard confusion metrics which is used to summarize the predictive performance of a classifier on test data. A single prediction by a classifier can have four outcomes as shown in Table-4.

TABLE IV
CONFUSION MATRIX

	Predicted *True*	Predicted *False*
Actual Class *True*	TP	FN
Actual Class *False*	FP	TN

In Table-4, *TP* represents as *True Positive*, which refer to the positive data that are correctly labeled by the classifier, *FN* represents as *False Negative*, which refer to the positive data that are mislabeled as negative. *FP* represents as *False Positive*, which refers to the negative data that are incorrectly labeled as positive and *TN* represents as *True Positive*, which are the negative data that are correctly labeled.

Using Table-4 matrix we can compute True Positive Rate (TPR) as.

$$TPR = \frac{TP}{(TP + FN)}$$

and, we compute False Positive Rate (FPR) as,

$$FPR = \frac{FP}{(FP + FN)}$$

For efficient intrusion detection TPR should be high and FPR should be low.

B. Classification Accuracy

Classification Accuracy (CA) is used to measure the classifier accuracy. Based on the confusion matrix we measure CA as,

$$CA = \frac{(TP + TN)}{(TP + FN + FP + TN)} \times 100$$

C. Root Mean Square Error

Root Mean Squared Error (RMSE) is used to measure mean absolute error rate. It is computed using the square root of the mean squared error and the resulting error. This is useful and allows for error measurement in the same magnitude to be quantity being predicted. It is measured as shown below and where, *d* is the number of record in test data, and *y* is the values in records.

$$RMSE = \frac{\sum_{i=1}^{a}(y_i - y_i')^2}{d}$$

A. Result Evaluation

On the reduced data set, we applied in the following classifier to measure performance measure using WEKA Tool.

- Naïve Bayes**:** It provides probabilistic knowledge through learning to a classifier for classification [8].
- Simple Cart**:** It is defined as classification and regression tree. It provides prior probability distribution mechanism to generate the regression tress for classification [4].
- J48**:** It is a derived version of C4.5 algorithm based on tree classifier and developed by Quinlan [4][11].
- NB Tree**:** Naïve Bayes Tree implements naïve classifier to build a structured tree, where each leaf node implements a decision tree. It provides an advantage of integration of naïve classifier and decision tree classifier [6].

A. *Classifier Accuracy Comparison*

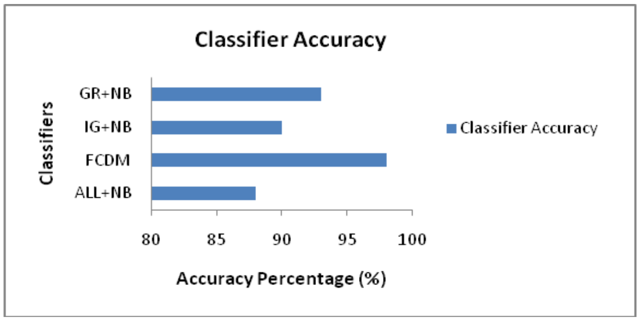

Fig.2 Classifier Accuracy Comparison using selected attributed sets in Table-3

B. True and False Positive Rate Comparison

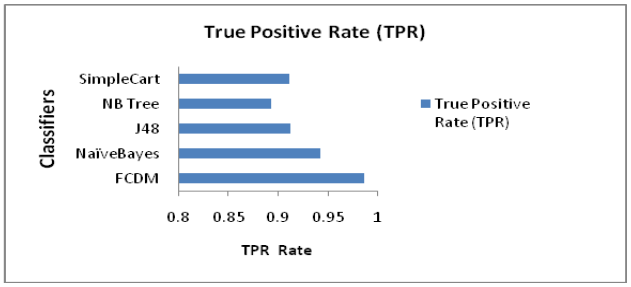

Fig.3 TPR Comparison using FCD Features selection

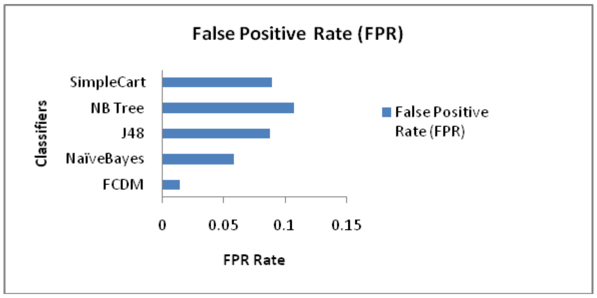

Fig.4 FPR Comparison using FCD Features selection

C. Root Mean Square Error Comparison

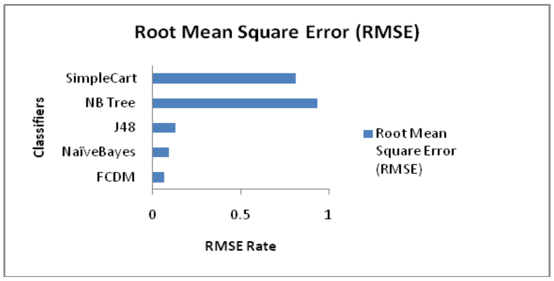

Fig.5 RMSE Comparison using FCD Features selection

The comparison results of classifiers in Figure-2,3,4 & 5 in relates to accuracy, TPR, FPR and RMSE shows that proposed FCDM achieved a better accuracy and high TPR and low FPR and RMSE.

VII. Conclusion

Intrusion detection is challenging issue in network log data collection. Naïve Bayesian classifiers prove a better efficiency in compare to other classifiers due to its high analyzing and auditing capacity over large datasets. A network log provides a huge collection of data and features for analysis. It requires an effective feature selection approach for better classification and detection which is a major impact on intrusion detection. We propose a feature covariance deviation approach for effectively reducing the features selection and modified the Naïve Bayesian algorithm to achieve high accuracy in intrusion detection. Experiment evaluation in compare with existing feature selection and classifiers shows an improvisation in classification by minimizing the false positive and root mean squired rate. The proposed work selects feature based on covariance deviation of all attack features of a class. The work can further improvise to evaluate individual attack features of a class for detail features variation and also in integration with other classifiers in the future works.

References

[1] M. Tavallaee, E. Bagheri, W. Lu and A.Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” Submitted to 2nd IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009.

[2] D. Ndumiyana, R. Gotora and H. Chikwiriro, “Data Mining Techniques in Intrusion Detection: Tightening Network Security”, International Journal of Engineering Research & Technology, Vol. 2 Issue 5, May, 2013

[3] H. Tribak , Blanca L. et.al.,, “ Statistical Analysis of Different Artificial Intelligent Techniques applied to Intrusion Detection System”, IEEE, 2012 http://dx.doi.org/10.1109/ICMCS.2012.6320275

[4] S. Thaseen and Ch. Aswani K, “An Analysis of Supervised Tree Based Classifiers for Intrusion Detection System”, International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME), IEEE, February 21-22 2013 http://dx.doi.org/10.1109/ICPRIME.2013.6496489

[5] M K. Asif, Talha A. K, et. al.,, “ Network Intrusion Detection and its Strategic Importance”, Business Engineering and Industrial Applications Colloquium(BEIAC), IEEE, 2013

[6] R. Kohavi, “Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid,” ser. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996, pp. 202–207.

[7] Zubair A. B, A S. Shaheen, and Radwan A, “An AODE-based Intrusion Detection System for Computer Networks,” pp. 28–35, IEEE 2011.

[8] Panda M. and M R Patra, “A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection”, First International Conference on Emerging Trends in Engineering and Technology, pp 504-507, IEEE, 2008 http://dx.doi.org/10.1109/ICETET.2008.80

[9] H Nguyen, K Franke and S Petrovi’c, “Improving Effectiveness of Intrusion Detection by Correlation Feature Selection,” International Conference on Availability, Reliability and Security, pp. 17–24, IEEE 2010. http://dx.doi.org/10.1109/ARES.2010.70

[10] Meng J, S Haikun, “The application on intrusion detection based on K-Means cluster algorithm,” International Forum on Information Technology and Application, 2009

[11] Mohammadreza E, Sara M, Fatimah S, L S Affendey, “Intrusion Detection Using Data Mining Techniques,” Proceedings Of IEEE International Conference on Information Retrieval & Knowledge Management,Exploring Invisible World, CAMP’10, pp.200-203,2010.

[12] Y. Li u, X. Yu, J.X. Huang, A.” An, Combining integrated sampling with SVM ensembles for learning from imbalanced datasets”, Information Processing &Management 47-617–631,2011

[13] S.-J. Horng, M.-Y. Su, Y.-H. Chen, T.-W. Kao, R.-J. Chen, J.-L. Lai, C.D. Perkasa,” A novel intrusion detection system based on hierarchical clustering and support vector machines”, Expert Systems with Applications,38-306–313.,2011

[14] Q. Zhang, G. Hu and W. Feng, “Design and performance evaluation of a machine learning-based method for intrusion detection”, in: Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed computing, in: Studies in Computational Intelligence, vol. 295, Springer,pp. 69–83. 2010. http://dx.doi.org/10.1007/978-3-642-13265-0_6

[15] J Zhang and M. Zulkernine, “Anomaly based Network Intrusion detection with unsupervised outlier detection,” School of Computing Queen’s University, Kingston, Ontario, Canada. IEEE International Conference ICC 2006, Volume-9, pp. 2388-2393, 11-15 June 2006.

[16] K Wankhade, S Patka and R Thools, "An Efficient Approach for Intrusion Detection Using Data Mining Methods", IEEE 2013.

[17] R Chitrakar and H Chuanhe, “Anomaly Detection using Support Vector Machine Classification with k-Medoids Clustering”, IEEE, 2012 http://dx.doi.org/10.1109/AHICI.2012.6408446

[18] Gary S, B Chen,“ Decision Tree Classifier for network intrusion detection with GA based feature selection,” University of Central Florida. ACM-SE 43, Proceedings of 43rd annual southeast regional Conference. Volume-2, ACM, 2005.

[19] Cuixiao Z, G Zhang, Shanshan S., “A mixed unsupervised clustering based Intrusion detection model,” Third International Conference on Genetic and Evolutionary Computing, 2009

[20] “Knowledge Discovery in Databases DARPA archive. Task Description”, KDDCUP-1999 DataSet, http://www.kdd.ics.uci.edu/databases/kddcup99/task.htm.

[21] NSL-KDD data set for network based intrusion detection systems. Available on: http://nsl.cs.unb.ca/NSL-KDD/, March 2009.

[22] Fatin N, Mohd Sabri, Norita Md Norwawi and K Seman, “Hybrid of Rough Set Theory and Artificial Immune Recognition System as a Solution to Decrease False Alarm Rate in Intrusion Detection System”, IEEE 2011

[23] Z.Muda, W Yassin, M.N. Sulaiman and N.I. Udzir, “Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification” 7th IEEE International Conference on IT in Asia , 2011.

B.V.RamNaresh Yadav is presently working as Assistant Professor in the Department of Computer Science and Engineering, JNTUH College of Engineering Nachupally, Karimnagar, Telangana State, India. He is a research scholar from the JNTUH University, Hyderabad. He has over 13 years of teaching experience. His areas of specializations include Network Security, Data Mining and Compiler Design.

Dr. B.Satya Narayana is presently working as a Professor in the department of Computer Science & Technology, SK University, Ananthapur, Andhra Pradesh State, India. He has published several Research papers in various National and International Journals. He is presently BOS Chairman for Dept. of CST in the same university. He has more than 20 years of teaching experience. His areas of specializations include Computer Networks, Data Mining and Artificial Intelligence.

Dr. D.Vasumati is presently working as a Professor in the department of Computer Science & Engineering, JNTUH CEH, Hyderabad, Telangana State, India. She has published several Research papers in various National and International Journals. She has more than 15 years of teaching experience. Her areas of specializations include Computer Networks, Data Mining and Big data.

http://dx.doi.org/10.15242/IIE.E0415046