ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

A Hybrid Intrusion Detection System Using K-Means Clustering and J48 Classification

[1] Navita Datta
[1] M.Tech Student, Department of Computer Science Engineering, DAVIET, Jalandhar .

Abstract: This paper is based on a hybrid intrusion detection system by using integrating K-means clustering and J48 classification. Firstly, the features are selected using correlation based feature selection, so that the number of attributes participating in detection of attacks can only be taken into concern and then it reduces the dimensionality of the attributes using Principal Component and Analysis. This algorithm works on the NSL-KDD dataset which is an improved version of the previously used KDD CUP'99 Dataset. Then we apply K-Means clustering over the obtained attributes and lastly we apply J48 classification for its evaluation. The proposed work has been fulfilled with an increase in accuracy and decrease in False Positive Rate.

I. INTRODUCTION

In our day to day life, security is becoming a big issue. Even if we talk of network then security is becoming a big hurdle, for the systems connected through internet. The networks have to be secured from a number of attacks like Denial of service (DoS), R2L, U2R and probe attacks. The various attacks related to network can be considered as an intrusion. Intrusion is characterized as "any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource". For controlling intrusion, intrusion discovery systems are utilized.

An IDS framework is a resistance framework, which recognizes undermining activities or attempts in a system.1,2IDS are ordered into 2 classes as indicated by the interruption discovery approach utilized: Misuse Detection Approach or Signature Approach and Anomaly Detection Approach used.(Anderson, 1995; Rhodes, Mahaffey, & Cannady, 2000; Tiwari, 2002).Misuse Detection Systems also known as Signature Detection System uses the stored information for the detection of intrusions i.e. these systems match the computer activity with the activities stored in their memory, any match of computer activity with the stored activity can be considered as an intrusion otherwise the activity can be considered as a normal activity. Essentially if there should be an occurrence of peculiarity identification approach, the interruptions are distinguished on the premise of standards that are connected over to the PC exercises to identify anything that goes amiss from the typical movement.(Kemmerer & Vigna,2002; Ingham,2003) These rules are based on the learning mechanism of anomaly detection approach to detect for the normal activities. The main advantage of misuse detection approach is the high detection rate and accuracy for all the known attacks whereas in case of anomaly detection approach, the principle advantage lies in identifying obscure and more muddled interruptions. The significant weakness of misuse detection approach is that it can detect only the known attacks stored in their database whereas in case of anomaly detection approach, the major shortcoming lies in its low discovery rate and high false alert rate.

As indicated by the assets they screen, IDS are ordered into two classes as: Host based IDS and Network based IDS. Host construct IDS are accessible with respect to the neighborhood have machine and assess the exercises and access to key servers. The Network based IDS systems distinguish and review the bundles going through the network.

This paper is based on a hybrid intrusion detection system (HIDS). The proposed system consolidates the positive parts of both the methodologies with a specific end goal to accomplish higher discovery rate, high exactness, low false alert rate and along these lines a raised level of execution. In this paper we likewise connected g clustering technique.
Clustering is a procedure by which we mark the information and dole out that information into different gatherings with the end goal that each gathering contains an accumulation of comparable items.

K-Means is one of the easiest unsupervised getting the hang of bunching calculations. It is a simple approach to arrange a given dataset through a specific number of k groups. In this technique first k focuses positions (C1,… …Ck) are instated. At that point every information point xi is allocated to the closest group focus Ci. The places of the K bunch focuses are recalculated over and over until the point that the group focuses never again change their positions i.e. until the point when the joining point is come to. The formatter should make these parts, joining the material criteria that take after.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Whatever is left of the paper is composed as takes after. Section 2 discusses the related works and Section 3 defines the proposed methodology. In Section 4, the experimental results are calculated and their comparison with the existing work is given. Finally, Section 5 deals with the conclusion of paper.

II. LITERATURE SURVEY

The proposed technique portrayed in segment 3 intends to accomplish high precision, high location rate and low or no false alert rate. It talks about constraints of past strategies including points of interest related to proposed strategy.

The Y-implies bunching calculation [9] has improved identification value and low false caution value. Be that as it may, it can't tackle ongoing irregularity identification, since it can't refresh the date set powerfully amid the procedure.

The significant points of K-implies [10] are its lightweight, quick iterative calculation that is straightforward and actualize. Be that as it may, the significant downsides are its affectability to beginning conditions, for example, the quantity of segments and the underlying centroids, and it is additionally touchy to anomalies and clamor.

A grouping calculation utilizes K-Means and SOM [11] conquers restriction of customary SOM that can't give exact bunching outcomes which additionally defeats hindrance of conventional K-implies that relies upon underlying quality and it is likewise hard to discover the a fitting focal point of the bunch.

A parallel bunching gathering calculation [12] shapes the groups all the more expediently to mass information. It likewise accomplishes high identification rate yet false alert value is low.

Half and half learning approach [13] utilizes K-implies bunching and gullible bayes arrangement conquers disadvantages of direct identification rate and high false alert value.

The blend variation from the norm area structure [14] which merges K-means and two classifiers: k-nearest neighbor and straightforward bayes vanquish limitations of high false ready value in surviving technique. To enhance Accuracy, identification value and diminish false alert value, this paper introduces a half breed approach for interruption recognition framework. Highlight choice aides in choosing essential and important components from the informational index and lessens the time required to handle the informational index. Clamor as well as exceptions on the dataset is decreased by using separating strategy. By utilizing separate and combine and the separation of every dot the quantity of group centroids and suitable beginning centroids are figured consequently. It defeats all disadvantage of basic K-implies calculation. As single bunching calculation is hard to acquire an immense viable location, grouping troupe is utilized for the viable recognizable proof of some known and obscure examples of assaults to accomplish rich exactness, discovery value and also low false caution rate.

IV. PROPOSED METHODOLOGY

In this paper we introduce another hybrid intrusion detection system, the propose framework consolidates the positive parts of both misuse detection approach and anomaly detection approach to achieve higher detection rate, accuracy, low false alarms etc. the research aims to produce better approach in the hybrid intrusion detection system then the existing approaches that will be based on data mining techniques such as machine learning algorithms, apriori algorithm and association rule mining.

In this thesis work we will be designing the hybrid intrusion detection system combining the both misuse and anomaly detection approaches which will be analyzing the NSL-KDD data set generated by the network log or training data. The input for the new work will be provided by the training data of NSL-KDD data set and is tested by the testing data set .the new system will monitor the network logs and find out all the possible attacks in the data set and provide high detection rate and accuracy.

1) ARCHITECTURE OF THE PROPOSED METHODOLOGY

The architecture of the proposed work mentioned in fig1. is as follows: First of all the proposed work takes as an input NSL-KDD Dataset. There are 41 features in the NSL-KDD data set. Since not all are required to classify the instances into attacks and normal instances. So the main objective of the proposed model is to reduce the number of features required for the classification.
Feature selection is the process of choosing a subset of relevant features or attributes required for the classification. It is one of the most important and frequently used techniques in data pre processing for data mining. The main objective of the feature selection is to remove the redundant and irrelevant features from the dataset that provide little or more information for classification of instances. Feature selection plays an important role for time consumption and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

classification accuracy. There are three types of feature selection methods .These are: 1) Best first search, 2) Genetic search and 3) Rank search. We have chosen the best first search method for our feature selection in correlation based feature selection strategy .It provide a new feature set Ff, which is our final feature set. In the best first approach, the space of the attribute subsets is searched using greedy hill climbing which is augmented with a back tracking facility. The number of consecutive non- improving nodes allowed can be set to control the level of back tracking done. Best first search can be started with an empty set of attributes and search forward or it may start with full set of attributes and search backward. It may also start at any point and search in both direction.

This feature set is then provided to the principle component analysis (PCA) which helps in the reduction of dimensionality of the chosen features. This reduces dimensionality further helps in the easy classification of chosen attributes. It is the method of identifying the patterns in data and expressing the data by defining their resemblance and deviation. To find the patterns in such high dimension are very hard to get and PCA considers as a right tool for analyzing such kind of data. The other primary favorable position of PCA is that you can pack the information by decreasing the quantity of measurements without much loss of data. PCA acts as takes after: 1) Get a few information, 2) subtract the mean, 3) figure the co-fluctuation network, 4) compute the Eigen vectors and Eigen value of co-change grid, 5) Choosing parts and framing a component vector, 6) driving the new informational collection.

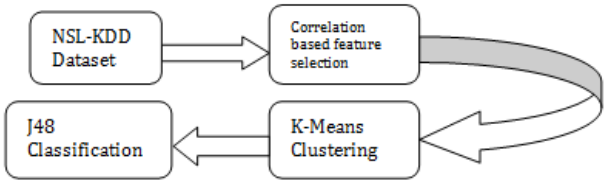

figure 1.Proposed System Architecture

Theorem 3 .1 Improved K-Means Clustering

Proof:

Randomly select initial centroids from this PCA data

Register the separation between every information point di (1 <= i<= n) to all the underlying centroids Cj (1 <= j<= k).

Repeat

For every information point di, locate the nearest centroid Cj and relegate di to bunch j.

Set ClusterId[i]=j. // j:Id of the closest cluster.

Set NearestDist[i]= d(di, cj).

For each cluster j (1 <= j <= k), recalculate the centroids.

For every information point di, process its separation from the centroid of the present closest group.

If

distance is not exactly or equivalent to the past cycle remove, the information point remains in a similar bunch. Else

For every centroid cj (1<=j<=k)register the separation d(di, cj).

End for;

until point that the joining criteria are met.

Theorem 3.2Improved J48 Algorithms

Proof:

Input Data:

1) D, a set of d training tuples;

2) k, the number of models in the outfit;

3) a learning scheme j48;

Output: A complex system, M∗.

Procedure:

For i = 1 to k do // create k models:

Create bootstrap sample, Di, by sampling D with weights;

Use Di to derive a model, Mi;

End for

To use the composite model on a tuple, X:

(a) if arrangement at that point

(b) let each of the k models order X and restore the greater part vote;

(c) if expectation at that point

(d) let each of the k models foresee an incentive for X and restore the normal anticipated esteem;

IV. EXPERIMENTAL EVALUATION AND RESULTS

The whole experiment was directed with assistance of JAVA Programming language, WEKA 3.6 machine learning tool and WEKA library capacities for feature selection techniques.

To recreate the exhibited thoughts, we utilize the NSL-KDD Dataset. The TCP dump crude information has been handled into association records, which are around five millions in which an informational index contains 24 assault sorts. Every one of these assaults lying under four principle classifications: DoS, U2R, and R2L, Probe as takes after. Typical joints are created with the help of catching day by day conduct, for example, downloading records or going to site page.

Denial of Service (DoS): The aggressor makes some figuring assets excessively occupied or memory assets too full to deal with honest to goodness asks for, or denies true blue clients request to a system. DoS assaults are characterized in light of

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

the administrations that the assailant makes inaccessible to other clients like apache2, arrive, mail, back, and so on.

Remote to Local (R2L): Various bundles are sent by those aggressors having no record on a remote machine which adventures few weakness to increase nearby accessibility of client on that system. It also incorporates send-letters, and Xlock.

User to Root (U2R): The assailant begins with access as a typical client on system which turns into a root client with the help of by misusing vulnerabilities to pick up the root access.
Probing: The assailant filters a system to gather data or to discover known vulnerabilities. An aggressor having a guide of systems and administrations which are accessible on the system can utilize the data to search for abuses.
With a specific end goal to assess the execution of this strategy we have utilized NSL-KDD informational index. To begin with we apply the connection based component determination calculation, and after that K-implies bunching calculation on the elements chose. From that point forward, we group the acquired information into Normal or Anomalous bunches by utilizing J48 classifier.

The importance of true positive (TP), true negative (TN), false positive (FP), false negative (FN) are characterized as takes after.

True positive (TP): The number of vindictive records which are accurately named intrusion.
True negative (TN): The number of honest to goodness records which are not delegated intrusion.
False positive (FP): The number of records which are inaccurately delegated assaults.
False negative (FN): The number of records which are inaccurately delegated true blue exercises.

The following parameters are used for the evaluation of our work and these parameters are described as below.

Precision:- Precision or Exactness is the division of recovered occasions that are important. Precision can been viewed as a measure of quality. In straightforward terms ,high precision implies that a calculation returned more important outcomes than insignificant ones. The formula for precision is defined in Eq(1).

$$\text{Precision}=\frac{TP}{TP+TN} \qquad \text{.....(1)}$$

The following table in figure 2depicts the precision calculation for the existing work and proposed work.

Precision	Existing Work	Proposed Work
10% Dataset	0.981	0.9996
15% Dataset	0.985	0.9952
20% Dataset	0.989	0.9969

figure 2: Precision Comparison

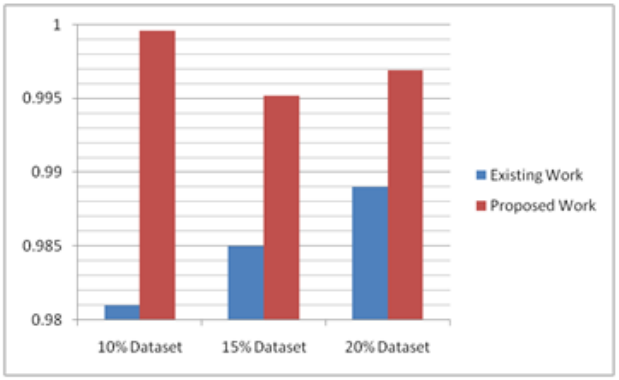

figure 3: Precision comparison b/w existing & proposed work

The above graph in figure 3 depicts that there is an increase in the precision parameter b/w the existing and the proposed work. This increase is high if the dataset size is very small as it takes into concern less number of instances whereas if we increase the size of the dataset i.e. if there is an increase in the number of instances taken into concern then the value of precision firstly decreases and then increases.

Recall:- It is the division of significant examples that are recovered. Recall is a measure of Completeness and Quantity. High Recall implies that a calculation returned the vast majority of the significant outcomes. The formula for recall is defined in Eq(2).

$$\text{Recall}=\frac{TP}{TP+FN} \qquad \text{.....(2)}$$

The following table in figure 4 depicts the recall calculation for the existing work and proposed work.

Recall	Existing Work	Proposed Work
10% Dataset	0.985	0.9983
15% Dataset	0.989	0.9942
20% Dataset	0.99	0.995

figure 4: Recall Comparison

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

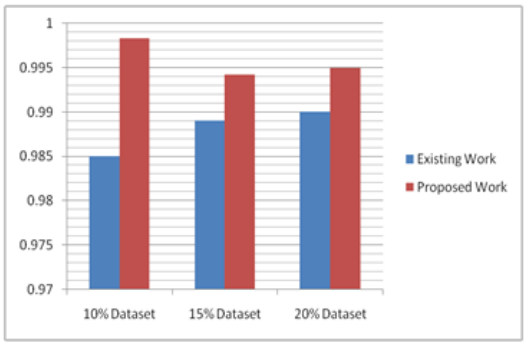

figure 5: Recall comparison b/w existing & proposed work

The above graph in figure 5 depicts that there is an increase in the recall parameter b/w the existing and the proposed work. This increase is high if the dataset size is very small as it takes into concern less number of instances whereas if we increase the size of the dataset i.e. if there is an increase in the number of instances taken into concern then the value of recall firstly decreases and then increases.

Accuracy:- Accuracy of classifier alludes to the capacity of classifier. It anticipates the class name effectively and the Accuracy of the indicator alludes to how well a given indicator can figure the estimation of anticipated property for another information. The formula for accuracy is defined in Eq(3).

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad \ldots\ldots\ldots\ldots(3)$$

The following table in figure 6 depicts the accuracy calculation for the existing work and proposed work.

Accuracy	Existing Work	Proposed Work
10% Dataset	98.2296%	99.8492%
15% Dataset	98.6292%	99.6242%
20% Dataset	98.8687%	99.7103%

figure 6: Accuracy Comparison

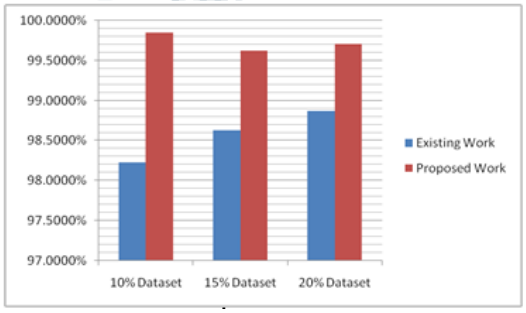

figure 7: Accuracy comparison b/w existing & proposed work

The above graph in figure 7 depicts that there is an increase in the accuracy parameter b/w the existing and the proposed work. This increase is high if the dataset size is very small as it takes into concern less number of instances whereas if we increase the size of the dataset i.e. if there is an increase in the number of instances taken into concern then the value of accuracy firstly decreases and then increases.

F-Measure:- The F-Measure is defined as a term which calculates the harmonic mean of the parameters precision and recall. The formula for F-Measure is defined in Eq(4).

$$\textbf{F-Measure} = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}} \quad \ldots\ldots\ldots\ldots\ldots\ldots(4)$$

The following table in figure 8 depicts the F-Measure calculation for the existing work and proposed work.

F-Measure	Existing Work	Proposed Work
10% Dataset	0.982	0.998
15% Dataset	0.986	0.996
20% Dataset	0.989	0.997

figure 8: F-Measure Comparison

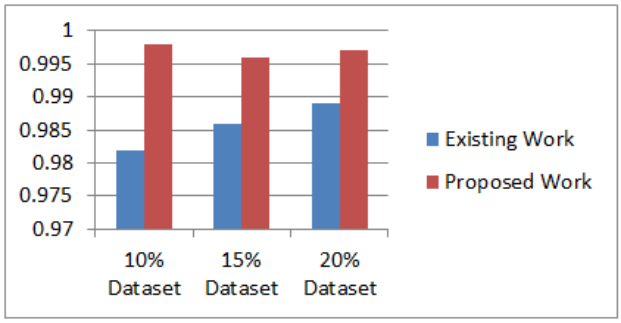

figure 9: F-Measure comparison b/w existing & proposed work

The above graph in figure 9 depicts that there is an increase in the F-Measure parameter b/w the existing and the proposed work. This increase is high if the dataset size is very small as it takes into concern less number of instances whereas if we increase the size of the dataset i.e. if there is an increase in the number of instances taken into concern then the value of F-Measure firstly decreases and then increases.

FP Rate:- The term false positive proportion, otherwise called the false alert proportion, for the most part alludes to the likelihood of dishonestly dismissing the invalid speculation for a specific test. The false positive rate is computed as the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

proportion between the quantity of negative occasions wrongly sorted as positive (false positives) and the aggregate number of genuine negative occasions (paying little heed to characterization). The false positive rate (or "false alert rate") more often than not alludes to the anticipation of the false positive proportion. The formula for FP-Rate is defined in Eq(5).

$$\text{FP Rate} = \frac{FP}{FP+TN} \qquad \text{.....(5)}$$

The following table in figure 10 depicts the FP Rate calculation for the existing work and proposed work.

FP Rate	Existing Work	Proposed Work
10% Dataset	0.017	0.0041
15% Dataset	0.013	0.0031
20% Dataset	0.012	0.0027

figure 10: FP Rate Comparison

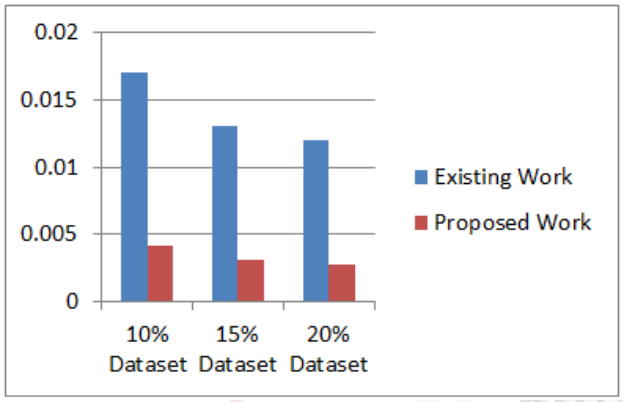

figure 11: FP Rate comparison b/w existing & proposed work

The above graph in figure 11 depicts that there is a decrease in the FP Rate parameter b/w the existing and the proposed work. This decrease is high if the dataset size is very small as it takes into concern less number of instances whereas if we increase the size of the dataset i.e. if there is an increase in the number of instances taken into concern then the value of FP Rate keeps on decreasing.

ROC Area: - The ROC bend is made by plotting the genuine positive rate (TPR) against the false positive rate (FPR) at different limit settings. FPR and TPR on x and y axes respectively are used for the characterization of ROC space that delineates relative exchange offs between genuine positive (benefits) and false positive (costs). Since TPR is identical to affectability and FPR is equivalent to 1 − specificity, the ROC diagram is now and then called the affectability versus (1 − specificity) plot. Every expectation outcome or example of disarray framework speaks to one point in the ROC space.

The following table in figure 12 depicts the ROC Area calculation for the existing work and proposed work.

ROC Area	Existing Work	Proposed Work
10% Dataset	0.998	1
15% Dataset	0.998	1
20% Dataset	0.999	1

figure 12: ROC Area Comparison

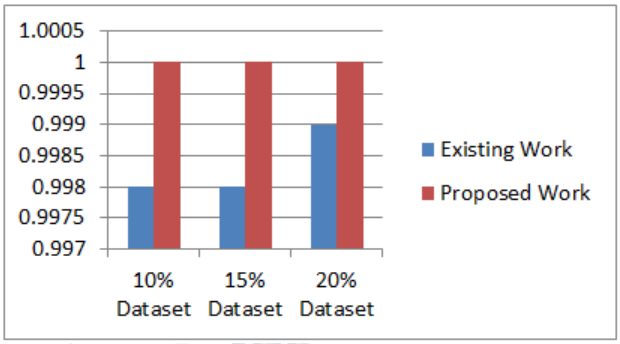

figure 13: ROC Area comparison b/w existing & proposed work

The above graph in figure 13 depicts that there is an increase in the ROC Area parameter b/w the existing and the proposed work. This increase is high if the dataset size is very small as it takes into concern less number of instances whereas if we increase the size of the dataset i.e. if there is an increase in the number of instances taken into concern then the value of firstly decreases and then increases. Since in the above graph the ROC Area value is constant for our proposed work because we have rounded off the value due to very little difference in various datasets.

CONCLUSION & FUTURE WORK

Based on past clustering technique, it is concluded that not any existing clustering method has high exactness, accuracy and review. Henceforth, a HIDS is proposed in this paper. The principle explore technique is grouping examination having target to accomplish high precision and low false positive value.

The vital qualities from the informational index are selected by Feature selection. Clamor and exceptions on the informational index are decreased by using filter method. Union aides and partition helps in figuring the k-number of bunch centroids. To get exact strategy for discovering introductory k grouping focuses, the interruption location show with bunching outfit is exhibited to accomplish high exactness and accuracy and review. Our future work can be led in the accompanying perspective: To develop newer forms of HIDS by using other techniques of feature selection.

- To try our strategy with genuine cloud information for encountering the constant impacts.

REFERENCES

[1] W. Lee and S. J. Stolfo. "Data mining approaches for intrusion detection," In Proceedings ofAntonio, TX, January 1998.

[2] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, "Intrusion Detection based on K-Means Clustering and OneR Classification", In Proceedings of 7th International Conference on Information Assurance and Security (IAS), IEEE, 2011, pp.192-197.

[3] Snehlata S. Dongre and Kapil K. Wankhade, "Intrusion Detection System Using New Ensemble Boosting Approach", In International Journal of Modeling and Optimization, Vol. 2, No. 4, August 2012, pp 488-492.

[4] Kapil Wankhade, MrudulaGudadhe, Prakash Prasad, "A New Data Mining Based Network Intrusion Detection Model", In Proceedings of International Conference on Computer and Communication Technology (ICCCT 2010), IEEE, 2010, pp.731-735.

[5] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification", In Proceedings of 7th International Conference on IT in Asia (CITA), IEEE, 2011.

[6] Shaik Akbar, Dr.K.Nageswara Rao, Dr.J.A.Chandulal, "Intrusion Detection System Methodologies Based on Data Analysis", In International Journal of Computer Applications (0975 – 8887) Volume 5– No.2, August 2010, pp.10-20.

[7] Deepthy K Denatious, Anita John, "Survey on Data Mining Techniques to Enhance Intrusion Detection", In Proceedings of International Conference on Computer Communication and Informatics (ICCCI - 2012), Jan. 10 – 12, 2012, Coimbatore, INDIA, IEEE, 2012.

[8] Kapil Wankhade, Sadia Patka, RavindraThool, "An Overview of Intrusion Detection Based on Data Mining Techniques", In Proceedings of 2013 International Conference on Communication Systems and Network Technologies, IEEE,2013, pp.626-629.

[9] Yu Guan and Ali A. Ghorbani, Nabil Belacel, "Y-Means: A Clustering Method For Intrusion Detection", In Proceedings of Canadian Conference on Electrical and Computer Engineering, Montreal, Quebec, Canada, May 4-7, 2003, IEEE, 2003, pp.1083-1086.

[10] Yang Zhong, Hirohumi Yamaki, Hiroki Takakura, "A Grid-Based Clustering for Low-Overhead Anomaly Intrusion Detection", IEEE, 2011, pp.17-24.

[11] A. Cardenas, J. Baras, and K. Seamon, "A framework for the evaluation of intrusion detection systems," in Proceedings of IEEE Symposium on Security and Privacy, (S&P), p. 15, 2006.

[12] G. Gu, P. Fogla, D. Dagon, W. Lee, and B. Skori´c, "Measuring intrusion detection capability: An information-theoretic approach," in Proceedings of ACM Symposium on Information, computer and communications security (ASIACCS06), pp. 90–101, ACM New York, NY, USA, 2006.

[13] Z. Muda, W. Yassin, M.N. Sulaiman, N.I. Udzir, "Intrusion Detection based on K-Means Clustering and Naïve Bayes Classification", In Proceedings of 7th International Conference on IT in Asia (CITA), IEEE, 2011.

[14] Hari Om, AritraKundu, "A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System", In Proceedings of 1st Int'l Conf. on Recent Advances in Information Technology (RAIT- 2012),IEEE, 2012.