INTERNATIONAL JOURNAL of RESEARCH –GRANTHAALAYAH
A knowledge Repository
Science

# AN IMPLEMENTATION OF IDS IN A HYBRID APPROACH AND KDD CUP DATASET

Mala Bharti Lodhi*[1], Prof. Vineet Richhariya[2], Prof. Mahesh Parmar[3]
[*1, 2, 3]CSE, LNCT Bhopal, India
*Correspondence Author: mala141086@yahoo.co.in

**Abstract:**
***Now in these days due to rapidly increasing network applications the data and privacy security in network is a key challenge. In order to provide effective and trustable security for network, intrusion detection systems are helpful. The presented study is based on the IDS system design for network based anomaly detection. Thus this system requires an efficient and appropriate classifier by which the detection rate of intrusions using KDD CPU dataset can be improved. Due to study there is various kind of data mining based, classification and pattern detection techniques are available. These techniques are promising for detecting network traffic pattern more accurately. On the other hand recently developed the hybrid models are providing more accurate classification. Thus a hybrid intrusion system is presented in this proposed work. That provides a significant solution even when the overall learning patterns are not available in database. Therefore, three different data mining algorithm is employed with system. Proposed system consists of K-mean clustering algorithm for finding the relationship among data in order to filter data instances. The implementation of the proposed classification system is performed using MATLAB environment and performance of designed classifier is evaluated. The obtained results from the simulation demonstrate after filtering steps. On the other hand the classification accuracy is adoptable with low number of training cycles with less time and space complexity.***

**Keywords:**
*IDS, Classification, KDD Cup 99's, MATLAB, Hybrid classification.*

## 1. INTRODUCTION

Now in these days the network communication is growing continuously and adopted rapidly. The network technology having a large number of applications, this is now used for banking applications, shopping and others. Thus a significant amount of sensitive and private data is traversing through these networks. Due to data traversing through untrusted network, the loss of security and data is an essential concern in the network technology. The presented study provides a detailed investigation of the security aspects and their flaws. Therefore, in this study intrusion detection systems are learned and a new concept of intrusion system design is presented using hybrid classification technique.

The proposed hybrid classification technique involves the implementation of cluster analysis, Genetic algorithm [6, 10] and the KNN algorithm for classifying the KDD cup dataset [11]. KDD cup data set includes the 41 attributes [12] and a class label, thus total 42 attributes are available for classification. In this data set the classes are divided into two major classes' normal and anomaly. The data set basically contains the basic network packets and their values for detection.

INTERNATIONAL JOURNAL of RESEARCH –GRANTHAALAYAH
A knowledge Repository
Science

The proposed study includes the techniques and methodologies by which the classification and recognition of the malicious packets are easily performed using the training of hybrid classifier. This section provides the basic overview of the proposed study work.

## 2. INTRUSION DETECTION SYSTEM

IDS systems are a kind of security filter designed using software or hardware configuration for protecting the network. Therefore, intrusion detection system (IDS) examines all inbound and outbound network activity such as packet transactions, user activities and recognizes suspicious patterns. These patterns are analyzed using any network administrator defined rules, predefined constrains for network or using machine learning algorithms. That may specify a network or system attack [5] from someone attempting to negotiate or break into a system. There are numerous ways to categorize IDS:

*MISUSE DETECTION VS. ANOMALY DETECTION:* in misuse detection, the IDS analyses the information it collects and compares it to huge databases of attack signatures. Fundamentally, the IDS look for a particular attack that has already been documented. Like a virus detection system, misuse detection software is only as better as the database of attack signatures that it utilizes to evaluate packets against. In anomaly detection [3], the system administrator describes the baseline, or normal, state of the networks traffic load, breakdown, protocol, and typical packet size. The anomaly [7] detector observes network segments to evaluate their state to the normal baseline and look for anomalies.

*NETWORK-BASED VS. HOST-BASED SYSTEMS:* in a network-based system, or NIDS [4], the specific packets flowing through a network are analyzed. The NIDS [9] can identify malicious packets that are designed to be overlooked by firewalls simplistic filtering rules. In a host-based system, the IDS examines at the activity on each individual computer or host.

*PASSIVE SYSTEM VS. REACTIVE SYSTEM:* in a passive system, the IDS detect a potential security breach, log the information and signal an alert. In a reactive system, the IDS respond to the doubtful activity by logging off a user or by reprogramming the firewall to obstruct network traffic from the supposed malicious source.

Though they both relate to network security, an IDS varies from a firewall in that a firewall looks out for intrusions in order to stop them from happening. The firewall confines the access between networks in order to thwart intrusion and does not signal an attack from inside the network. An IDS calculates a supposed intrusion once it has taken place and signals an alarm. An IDS also watches for attacks that originate from within a system.

Working with all kind of system and is much complex, time consuming and expensive work for us, thus we work with first kind of system and modify it to make the IDS very powerful and attain better performance.

An Anomaly-Based Intrusion Detection System [1] is a system for detecting computer intrusions and misuse by monitoring system activity and classifying it as either normal or anomalous. The

INTERNATIONAL JOURNAL of RESEARCH –GRANTHAALAYAH
A knowledge Repository
Science

classification will identify any kind of misuse that falls out of usual system operation, which is depends on rules or heuristics, rather than patterns or signatures. It is as opposite to signature based systems which can only identify attacks for which a signature has previously been created.

In order to verify what attack traffic is, the system must be taught to recognize normal system activity. This can be accomplished in several ways, most often with artificial intelligence type techniques [2]. Systems using neural networks have been used to great effect. One more method is to describe what normal usage of the system includes using a strict mathematical model, and flag any deviation from this as an attack. This is termed as strict anomaly detection.

## 3. SYSTEM ARCHITECTURE

This section describes the overall description of the proposed architecture in a data process flow manner.

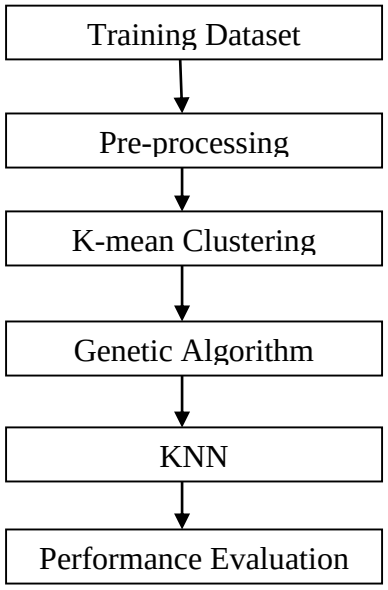


Figure 1 system architecture

The figure 1 system architecture describes the processing of the KDD cup dataset and their classification technique. Therefore first the KDD cup training dataset is pre-processed and attributes are converted into numerical data. Than after the K-mean traditional clustering algorithm is applied on the dataset. This produces the clusters of data. These clusters having two groups of data. The processing of k-mean clustering is given using below given table. During the pre-processing following variables are utilized:

INTERNATIONAL JOURNAL of RESEARCH –GRANTHAALAYAH
A knowledge Repository
Science

a. Number of instances $I_n$
b. Number of attributes $A_n$
c. Number of class labels $C_n$
d. Data set $D_{m,n}$

| |
|---|
| Input: $I_n, A_n, C_n, D_{m,n}$ |
| Output: $NewD_{m,n}$ |
| Process:<br>a. Read all instances from $D_{m,n}$<br>b. For each instances of $I_n$<br>c. [indx, labels] = K-mean $(D_{m,n},\ C_n, "ecludian\ distance")$<br>d. End for<br>e. Create a new Data array using new index and their class labels<br>f. $NewD_{m,n}$ |

Table 1 K-mean clustering

After cluster formation the data is processed using the genetic algorithm. The genetic algorithm [1,8, 10] work in four different steps

*1. POPULATION GENERATION***:** in this phase the data is evaluated again and for all the attributes the unique values are identified. Using this unique attribute values population generation is performed. The generated population is processed using the below given manner.

*2. SELECTION:* Every iteration results a new generation of possible solutions (individuals) for a given search issue. In first, a number of sequences known as population of possible solution are created as initialization. Every individual in population is treated by encoding them into a chromosome to enhance using genetic operators. In next step chromosomes are assessed, the individual is produced by its string description or chromosome, and then its performance in relation to the target response is computed. That is required to obtain how fit this individual is in relation to the other in the population. According to individual's fitness, a selection process selects the best pairs for the genetic modification. This selection technique is responsible to ensure the survival of the fittest individuals.

*3. CROSSOVER:* Crossover one of the genetic operators used to recombine the population's genetic material. It takes two chromosomes and swaps part of their genetic information to produce

INTERNATIONAL JOURNAL of RESEARCH –GRANTHAALAYAH
A knowledge Repository
Science

new chromosome. As figure shows after the crossover has been arbitrarily selected, segments of the parent's chromosome sequences.

*4. MUTATION:* The mutation operator provides new kind of genetic structures by randomly altering some of building blocks. Meanwhile the amendment is totally random and therefore not associated with previous genetic structures available in the population, "Mutation" it creates different structure associated to other segments of the search space. As shown in figure the mutation is implemented using altering an arbitrary bit from a chromosome sequence.

After finding the appropriate sequences in first round the KNN classification is applied to the generated sequences. And the following outcomes are evaluated on the test set.

| | |
|---|---|
| Normal | 12437 |
| U2R | 10291 |
| R2L | 10938 |
| DOS | 14028 |
| Probe | 12473 |
| Total | 60167 |

Table 2 test dataset

After applying the classification the following outcomes are observed.

| | | |
|---|---|---|
| Normal | 12349 | 88 |
| U2R | 10250 | 41 |
| R2L | 10905 | 33 |
| DOS | 14000 | 28 |
| Probe | 12448 | 25 |
| Total | 59952 | 215 |

Table 3 detection rate

According to the detection table 3 the overall performance of the system in terms of accuracy is 99.64% and the error rate of the system is .36%.

## 4. RESULTS

*A. SYSTEM PERFORMANCE*

INTERNATIONAL JOURNAL of RESEARCH –GRANTHAALAYAH
A knowledge Repository
Science

This section provides the complexity of classification according to the input training set. The evaluation of the performance is given in terms of memory and training time of implemented classifiers.

*MEMORY CONSUMPTION*

The amount of main memory consumed during algorithm processing is known as memory consumption of space complexity. The proposed classification performance is given with increasing size of dataset instances and their respective memory values.

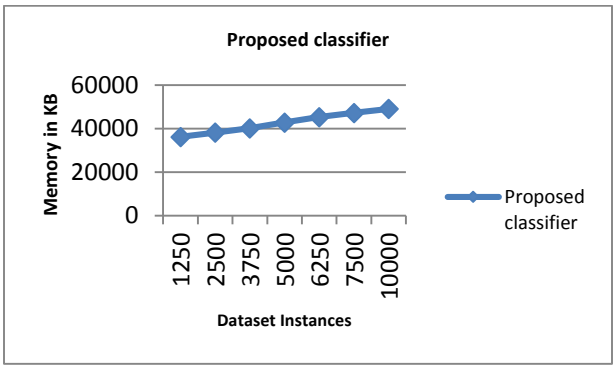


Figure 4.1 memory consumption

The evaluated memory consumption is provided using figure 4.1, in this diagram X axis provides the information about the dataset size in terms of data instances and respective memory consumption in terms of KB is given using Y axis. According to the evaluated results the obtained memory consumption is increases as the size of data set increase.

*TIME COMPLEXITY*

The amount of time required to perform classification task is known as the time complexity of the system. In this system three different classifiers are implemented one after another therefore the complexity is computed in terms of seconds with increasing size of data instances.

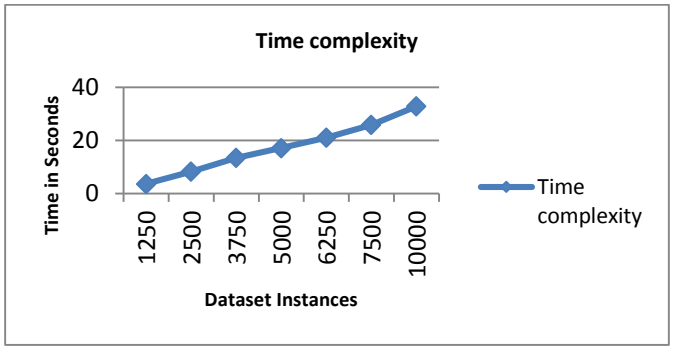


Figure 4.2 time complexity

INTERNATIONAL JOURNAL of RESEARCH –GRANTHAALAYAH
A knowledge Repository
Science

The time consumption of the proposed classification model is given using figure 4.2 in this diagram the amount of time is given in Y axis and the X axis demonstrate the size of dataset. According to the evaluated performance the classifier needs more time as the size of data is increases. Thus that can be conclude time complexity of the proposed classifier system is optimal and consumes an adoptable amount of time.

*B. IDS PERFORMANCE*

In this section the performance of IDS system for detection of malicious pattern, thus the obtained performance of the system is given in terms of accuracy and error rate.

*ACCURACY*

The amount of data correctly classified over the given input patterns is known as accuracy of the system. That can also derive using the formula given:

$$Accuracy = \frac{total\ correctly\ classified\ samples}{total\ input\ samples} X100$$

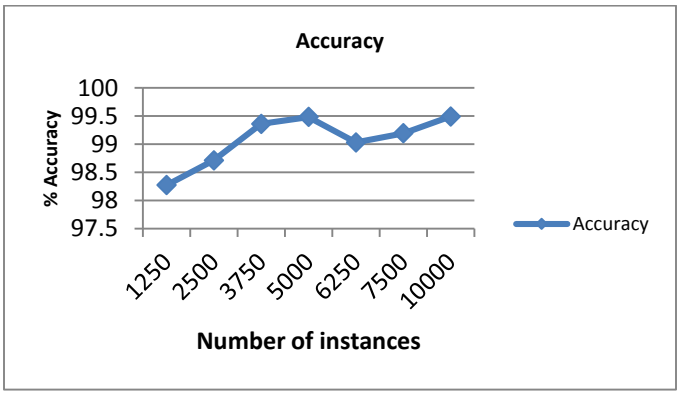


Figure 4.3 accuracy of system

The accuracy of the system is given using figure 4.3, in this diagram the percentage accuracy of the system is given in Y axis and the number of dataset instances are given using X axis. According to the evaluated results the performance of classification is improved if the class distribution and all the attributes are contains the significant information otherwise the performance of classification is decreases as given for 6250 instances. Thus the accuracy of the system is depends upon the data provided as input for learning.

*ERROR RATE*

The error rate provides the information about the misclassified data over the given samples to classify, in this scheme the N-cross validation processes used for calculating the accuracy and error rate. The obtained error rate can be calculated using the formula:

INTERNATIONAL JOURNAL of RESEARCH –GRANTHAALAYAH
A knowledge Repository
Science

$$error\ rate = \frac{total\ incorrectly\ classified\ samples}{total\ samples\ given\ for\ classification} X100$$

Or

$$error\ rate = 100 - accuracy$$


Figure 4.4 error rate

The obtained error rate from the classification system is given using figure 4.4 in this diagram the Y axis demonstrate the percentage error rate of the system and the X axis shows the number of instances in dataset.

## 5. CONCLUSIONS

Now in these days the communication system is affecting applications and use of applications, in this context security in this domain is a primary aspect in communication network. The proposed study is an investigation of IDS (intrusion detection system) and their design concept. For that purpose an intrusion detection system is developed using the analysis of KDD CUP 99's dataset. In this intrusion detection system the main focus is given over classification and performance improvement of classifiers. Therefore, different algorithms are applied for filtering the data set features.

The proposed IDS system utilizes the K-mean clustering algorithm, Bayesian classification algorithm and finally the back propagation neural network. The implementation of the desired system is performed using MATLAB IDE. And using the confusion matrix the performance is evaluated.

## *6. REFERENCES*

*[1] Hari Om, AritraKundu, "A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System", 1st Int'l Conf. on Recent Advances in Information Technology RAIT-2012, 978-1-4577-0697-4/12/$26.00 ©2012 IEEE*

INTERNATIONAL JOURNAL of RESEARCH –GRANTHAALAYAH
A knowledge Repository
Science

[2] *Gang Wang, Jinxing Hao, Jian Ma, Lihua Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering", 0957-4174, 2010 Elsevier Ltd. doi:10.1016/j.eswa.2010.02.102*

[3] *ManoranjanPradhan, Sateesh Kumar Pradhan, Sudhir Kumar Sahu, "Anomaly Detection using ArtificialNeural Network", International Journal of Engineering Sciences & Emerging Technologies, April 2012, ISSN: 2231 – 6604 Volume 2, Issue 1, pp: 29-36 ©IJESET*

[4] *Reyadh Shaker Naoum, NamhAbdulaAbid and ZainabNamh Al-Sultani, "An Enhanced Resilient Back propagation Artificial Neural Network for Intrusion Detection System", IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.3, March 2012, 11*

[5] *Vladimir Bukhtoyarov, Eugene Semenkin, "Neural Networks Ensemble Approach for Detecting Attacks in Computer Networks", WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia*

[6] *Xiao Hang Yao, "A Network Intrusion Detection Approach combined with Genetic Algorithm and Back Propagation Neural Network", 2010 International Conference on E-Health Networking, Digital Ecosystems and Technologies, 978-1-4244-5517-1/10/$26.00 ©201O IEEE*

[7] *Sufyan T. Faraj, Al-Janabi and HadeelAmjedSaeed, "A Neural Network Based Anomaly Intrusion Detection System", 2011 Developments in E- systems Engineering, 978-0-7695-4593-6/11 $26.00 © 2011 IEEE, DOI 10.1109/DeSE.2011.19*

[8] *XiaoHangYyao "A Network Intrusion Detection Approach combined with Genetic Algorithm and Back Propagation Neural Network" 2010 International Conference on E-Health Networking, Digital Ecosystems and Technologies*

[9] *Qinglei Zhang" Network Intrusion Detection by Support Vectors and Ant Colony" Proceedings of the 2009 International Workshop on Information Security and Application (IWISA 2009)Qingdao, China, November 21-22, 2009*

[10] *M. Sadiq Ali Khan" Rule based Network Intrusion Detection using Genetic Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 18– No.8, March 2011*

[11] *Prof. Mrs. N. S. Chandolikar" Efficient Algorithm for Intrusion Attack Classification by Analyzing KDD Cup 99" 2012 IEEE*

***[12]*** *Noreen Kausar "An Approach towards Intrusion Detection using PCA Feature Subsets and SVM" 2012 lEEE Universiti Teknologi PETRONAS, Bandar Seri Iskandar, 31750 Tronoh, Perak, Malaysia 2012 International Conference on Computer & Information Science (ICCIS)*