New Hybrid Intrusion Detection System based on Data Mining Technique to Enhanced Performance

Abhas Agrawal[1] Tripurari Pujan Pratap Singh[2] Kamal Niwaria[3]
[1]M. Tech Scholar [2]PhD scholar [3]Associate Professor
[1,2,3]Department of Electronic Engineering
[1,3]RKDF College RGPV MP India [2]Aisect University Bhopal M.P. India

Abstract— Intrusion Detection Systems (IDSs) is an efficient defense technique against network attacks as well host attacks since they allow network/host administrator to detect any type policy violations. However, Standard IDS are vulnerable and they are not reliable to novel and original malicious attacks. Also, it is very inefficient to analyze from a big amount of data such as possibility logs. Moreover, there are high false positives and false negatives for the common OSs. There are many other techniques which can help to improve the quality and results of IDS in which data mining one of them where it has been popularly recognized/identify as an important way to mine useful information from big amount of data which is noisy, and random. Integration of various data mining techniques with IDS to improve efficiency is the motive of proposed research. Proposed research is combining three data mining technique to reduce over head and improve execution efficiency in intrusion detection system (IDS). The Proposed research that ensembles clustering (Hierarchical) and two classifications (C5.0, CHAID) approaches. Proposed IDS execute on the standard KDD'99 (knowledge Discovery and Data Mining) Data set; this data set is used for measuring the performance of intrusion detection systems. Proposed system can detect the intrusions and classify them into four categories: Probe, Denial of Service (DoS), U2R (User to Root), and R2L (Remote to Local). A presented experiment results is carried out to the performance of the proposed IDS using KDD 99' dataset. Its shows that the proposed IDS performed better in term of accuracy, and efficiency.

Key words: Internet; Intrusion Detection; Data Mining; Clustering, Classification, Data Preprocessing

I. INTRODUCTION

Information security technology is an essential component for protecting public and private computing infrastructures. With the widespread utilization of information technology applications, organizations are becoming more aware of the security threats to their resources. No matter how strict the security policies and mechanisms are, more organizations are becoming susceptible to a wide range of security breaches against their electronic resources. Network-intrusion detection is an essential defense mechanism against security threats, which have been increasing in rate lately. It is defined as a special form of cyber threat analysis to identify malicious actions that could affect the integrity, confidentiality, and availability of information resources. Data mining-based intrusion-detection mechanisms are extremely useful in discovering security breaches. An intrusion detection system (IDS) is a component of the computer and information security framework. Its main goal is to differentiate between normal activities of the system and behavior that can be classified as suspicious or intrusive [11]. IDS's are needed because of the large number of incidents reported increases every year and the attack techniques are always improving. IDS approaches can be divided into two main categories: misuse or anomaly detection [12]. The misuse detection approach assumes that an intrusion can be detected by matching the current activity with a set of intrusive patterns. Examples of misuse detection include expert systems, keystroke monitoring, and state transition analysis. Anomaly detection systems assume that an intrusion should deviate the system behavior from its normal pattern. This approach can be implemented using statistical methods, neural networks, predictive pattern generation and association rules among others techniques. In this research using naïve byes classification with clustering data mining techniques to extract patterns that represent normal behavior for intrusion detection. This research is describing a variety of modifications that will have made to the data mining algorithms in order to improve accuracy and efficiency. Using sets of naïve byes classification rules that are mined from network audit data as models of "normal behavior." To detect anomalous behavior, it will generate naïve byes classification probability with clustering followed from new audit data and evaluate the similarity with sets mined from "normal" data. If the similarity values are below a threshold value it will show abnormality or normality [12].

II. PROPOSED WORK

This Chapter is going to be present general idea on a new proposed concept for intrusion detection system which will enhance efficiency as compare existing intrusion detection system. The proposed concept is using data mining techniques. Data mining techniques have been successfully applied in many different fields including marketing, manufacturing, process control, fraud detection, and network management. Over the past five years, a growing number of research techniques have applied data mining to various problems in intrusion detection. In this will apply to data mining for anomaly detection field of intrusion detection. Presently, it is unfeasible for several computer systems to affirm security to network intrusions with computers progressively getting connected to public accessible networks (e.g., the Internet). In view of the fact that there is no ideal solution to avoid intrusions from event, it is very significant to detect them at the initial moment of happening and take necessary actions for reducing the likely damage. One approach to handle suspicious behaviors inside a network is an intrusion detection system (IDS). For intrusion detection, a wide variety of techniques have been applied specifically, data mining techniques, artificial intelligence technique and soft computing techniques. Most of the data mining techniques like association rule mining, clustering and classification have been applied on intrusion

All rights reserved by *www.ijsrd.com*

detection, where classification and pattern mining is an important technique.

A. Proposed Concept:

Here proposed concept are going to be present general idea as showing in figure 4.1 for intrusion detection system which will enhance efficiency as compare existing intrusion detection system. The proposed concept is using data mining techniques. In this clustering and classification data mining technique has applied for anomaly detection field of intrusion detection. Anomaly learning approaches are able to detect attacks with high accuracy and to achieve high detection rates. However, the rate of false alarm using anomaly approach is equally high. In order to maintain the high accuracy and detection rate while at the same time to lower grade the false alarm rate, the proposed technique is the combination of three learning techniques. For the first stage in the proposed technique, this grouped similar data instances based on their behaviors by utilizing a hierarchical clustering as a pre-classification component. Next, using C5.0 classifier this classified the resulting clusters into attack classes as a final classification task. This found that data that has been misclassified during the earlier stage may be correctly classified in the subsequent classification stage. At last CHAID classification is applied. Following is the proposed IDS which divided into following module:

1) Database Creation (Suggested Technique)
- Download and Rearranged KDD 99'
- Data Formation and Re-Processing of KDD 99' (Training and Testing Data Set Prepration)

2) Data mining Techniques
- Cluster Technique
  - Hierarchical Clustering
- Classification
  - C5.0
  - CHAID

3) Proposed System
- K-Mean Clustering
- K-Mean Clustering with Naïve Bayse classification
- K-Mean with Naïve Bayse classification and Decision Table Majority Rule Based Approach
- Hierarchical Clustering
- Hierarchical Clustering with C5.0 classification
- Hierarchical Clustering with C5.0 classification and CHAID Classification

4) Performance
- Time Analysis
- Memory Analysis
- CUP Analysis

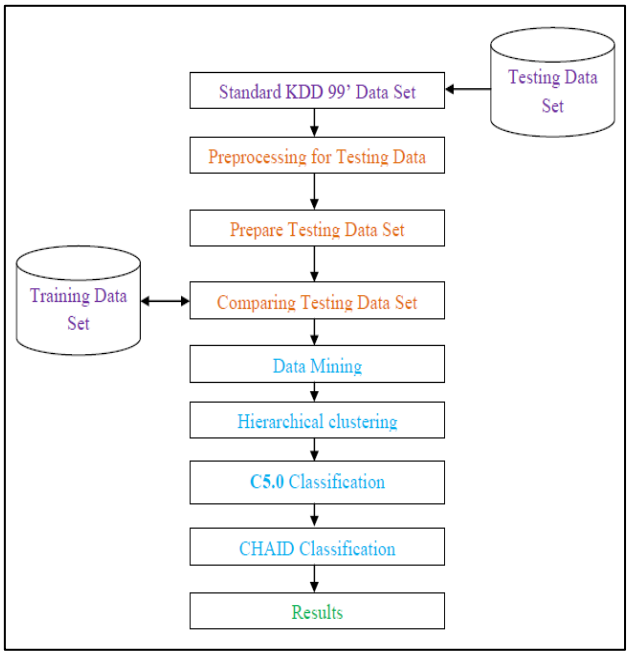


Fig. 4.1: Block Diagram of Proposed Concept

Proposed IDS Architecture: In the proposed work, outline a data mining approaches for designing intrusion detection models. The Basic idea behind this is that apply various data mining technique in single to audit data to compute intrusion detection models, as per the observation of the behavior in the data. In the proposed work are the combining three most useable data mining techniques into single concept and presenting architecture shown in figure 4.2. In proposed technique, use Hierarchical clustering, C5.0 algorithm and CHAID approach. First apply the hierarchical algorithm to the given dataset to split the data records into normal cluster and anomalous clusters. It specifies the number of clusters as five to the hierarchical and clusters the records in the dataset into normal cluster and anomalous clusters. The anomalous clusters are U2R, R2L, PROBE, and DoS. The records are labeled with the cluster indices. Then, divide the data set into two parts. One part is used for training and the other one is used for evaluation. In training phase, apply the labeled records to the C5.0 for training purpose. The C5.0 classifier is trained with the labeled records. Then, apply the rest of unlabeled records to the C5.0 for classification. The C5.0 classifier will classify the unlabelled record into normal and aberrant clusters. Finally apply CHAID which is also the classifier that is doing exact match of each attribute values all to gather and thus removes the strong independence assumption. The Proposed work consists of clustering, classification where proposed architecture as shown in figure 4.2.

All rights reserved by *www.ijsrd.com*

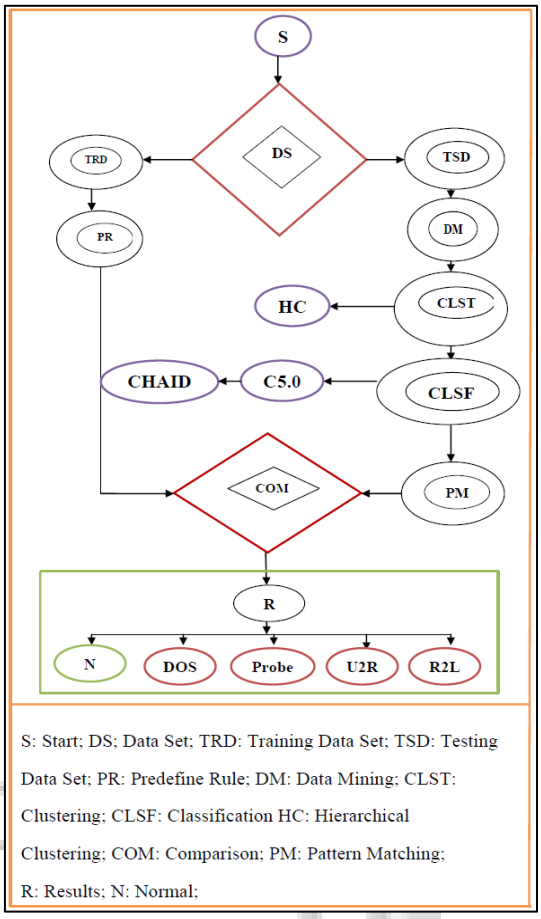


Fig. 4.2: Architecture of the Proposed IDS

B. Proposed Algorithm

Input: Dataset KDD, a sample K, Normal Cluster NC, Abnormal cluster AC, c is the number of clusters and d is the distance between them, ch1,ch2,ch3,ch4,ch5 are Nodes i1,i2,i3,i4 are the category

Output: K is abnormal or normal

C. Algorithm Hybrid

1) First apply Hierarchical clustering

1) Firstly load data into a root cluster and we start with one cluster and successively split clusters to produce others, more and more samples are clustered together in a hierarchical manner.
2) For Every data point:
3) Find out the distance from the data point to every cluster.

Begin

Initialize c; c' = n; $D_i = \{x_i\}$; i = 1,...,n

Do

c' = c' - 1

4) Find nearest clusters D_i and D_j
5) Merge D_i and D_j

Until c = c'

Return c clusters

End

6) To find the nearest clusters in step 4, the following clustering criterion function is used:
7) $d_{min}(D_i, D_j) = \min \| x - x' \|$, where $x \in D_i$ and $x' \in D_j$
8) The merging of the two clusters in step 6 simply corresponds to adding an edge between the nearest pair of nodes in D_i and D_j. Also, if instead of terminating after a predetermined number of clusters have been obtained; it is possible to set the termination criteria to stop when the distance between nearest clusters exceeds a predetermined threshold.

2) Apply C5.0 Classification

1) For each Clusters C in KKD_i in test data do

If C is i1
Ch1=c
Else
If C is i2
Ch2=c
Else
If C is i3
Ch3=c
Else
If C is i4
Ch4=c
Else
Ch5=c
until end of deta set

2) Collect data from dataset in the form of Normal/Abnormal and apply those data to the CHAID Decision Table Majority rule based approach and build condition for the action like training/testing normal data set D.

3) CHAID

1) Preparing predictors. The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations. For categorical predictors, the categories (classes) are "naturally" defined.

If (c is not equal to ch1,ch2,ch3,ch4)
Then
c is Normal
Otherwise
c is abnormal

III. RESULTS ANALYSIS

For experiment use a laptop Pentium® Dual-Core CPU T4400 @2.20Ghz and 32-bit operating system, in which performance data is collected. In the experiments, the laptop executes fixed record data sets (182679). Several performance metrics are collected:

– Execution time
– CPU Utilization time
– Memory Utilization

The execution time is considered the time that an algorithm takes to produce results. Execution time is used to calculate the throughput of an algorithm. It indicates the speed of algorithm. The memory deals with the amount of memory space it takes for the whole process of Intrusion Detection System. The CPU Utilization is the time that a CPU is committed only to the particular process of calculations. It calculates the load of the CPU. The more

All rights reserved by www.ijsrd.com

CPU time is used in the execution process, the higher is the load of the CPU. During Results evolution we have use the KDD99 cup data set [22, 23 & 24] for training and testing [1] which is shown in table 1 and 2. In 1998 DARPA intrusion detection evaluation program was set up to acquire raw TCP/IP dump data [21 & 22] for a LAN by MIT Lincoln lab to compare the performance of various intrusion detection methods [5 & 6]. In KDD-99 data set each record is consists of a set of features, some of which are either discrete or persistent. The qualitative values are labels without an order which could be symbolic or numeric values e.g. the value of feature protocol type is one among the symbols {icmp, tcp, udp}. The numeric value of the feature logged in is 0 or 1 to represent whether the user has successfully logged in or not. For the quantitative attributes, the data are characterized by numeric values within a finite interval. Example can be the duration. Since the feature selection is applicable only to the discrete attributes, not to the continuous ones, the continuous features need be converted to discrete ones prior to the feature selection analysis. In order to evaluate the performance of this method I have used KDD99 data set [27]. First apply K-means clustering algorithm on the features selected. After that, we classify the obtained data into Normal or Anomalous clusters by using the Hybrid classifier which is the combination of (K-nearest and Decision Table). In these experimental results compare packet performance, time-consuming, memory utilization and CPU utilization of known algorithm on fixed size of record sets. During processing, the record sets are coming from data base, table 1 is producing training data set and table 2 is producing testing data set. For evaluation mode, there are two parameters: the number of evaluated record set and the size of evaluated record set, where the number of evaluated record sets is the number of record set that are generated randomly and the size of evaluated record sets can be chosen from database. In this mode, n cycles (that is, the number of the evaluated record sets) executed. In each cycle, record sets are respectively executed by proposed technique. Finally, the outputs of the Proposed evaluation system are packet performance, execution time, and the execution time is measured in seconds. Actually, for an algorithm, the time-consuming of execution not only depends on the algorithm's complexity, but also the size of record sets. The evaluated results are illustrated as in Table 3 – 5.

Attacks Type	Training Example
Normal	170737
Remote to User	2331
Probe	7301
Denial of service	2065
User to Root	245
Total examples	182679

Table 1: Number of Example used in Training Data Taken from KDD99 Data Set

Attacks Type	Testing Example
Normal	78932
Remote to User	1015
Probe	4154
Denial of service	885
User to Root	145
Total examples	85131

Table 2: Number of Example used in Testing Data Taken from KDD99 Data Set

The execution time is considered the time that an algorithm takes to produce results. Execution time is used to calculate the throughput of an algorithm. It indicates the speed of algorithm. Table 3 is showing the execution time of proposed technique on 85131 testing data set.

Data Volume	Proposed IDS Technique(Hierarchical + C5.0+ CHAID)
	Execution Time in Millisecond (Approx)
85131	499

Table 3: Comparison of Execution Time on 85131 Data Volume

The memory deals with the amount of memory space it takes for the whole process of Intrusion Detection System. Table 4 is showing the memory utilization of proposed technique on 85131 testing data set.

Name	Total Available Memory	Total Memory Consumption	Memory Utilization in %
Proposed IDS Technique(Hierarchical + C5.0+ CHAID)	174568	12457	54

Table 4: Memory Utilization

The CPU Utilization is the time that a CPU is committed only to the particular process of calculations. It reflects the load of the CPU. The more CPU time is used in the execution process, the higher is the load of the CPU. Table 5 is showing the CPU utilization of proposed technique on 85131 testing data set.

Name	CPU Utilization in %
	(Approx)
Proposed IDS Technique(Hierarchical + C5.0+ CHAID)	60%

Table 5: of CPU Utilization

Here graph-1 is drawing form Table-3 to reveal it. In this graph, execution time is showing where the evaluated mode is fixed size of record sets ranging from 85131 approx testing record sets.

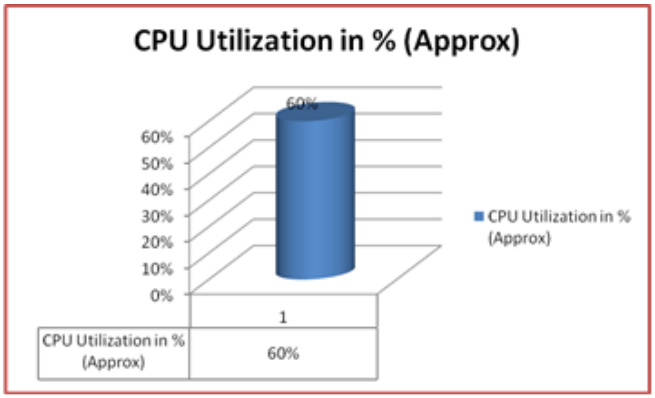
Graph 1:- Execution Time vs User Load of proposed technique on 85131 testing data set

Here graph-2 is drawing forms Table-4 to reveal it. In this graph memory utilization is showing where the evaluated mode is fixed size of record sets (85131).

Graph 2:- Memory Utilization of proposed technique and on 85131 testing data set

Here graph-3 is drawing forms Table-5 to reveal it. In this graph CPU utilization is showing where the evaluated mode is fixed size of record sets (85131).

All rights reserved by *www.ijsrd.com*



Graph 3: CPU Utilization of proposed technique and existing technique on 85131 testing data set

Experimental results for this compassion point are shown Table 3 to 5 at execution stage. The results show the superiority of proposed technique in terms of the processing time, Memory Utilization and CPU Utilization. Some typical results obtained by the evaluation system can be found in Tables (3, 4 & 5) and Graphs (1, 2 & 3). The results illustrated in Table 3 shows that proposed technique is in different/fixed record sets. Finally, it is not difficult to find that, in contrast with these Tables, the larger the data record sets, the bigger execution time is. Besides, in contrast with these Tables, it is not difficult to find that the increasing data length can lead to the significant increment of execution time as well as memory utilization and CPU Utilization. Generally speaking, the time-consuming of known algorithm usually depends on the size of record sets of.

A. Strength of the Proposed System

- Proposed Hybrid technique is producing good performance then comparing technique to find normal packet performance.
- Proposed hybrid technique having low response time than comparing technique.
- Proposed hybrid technique using low memory space during execution than the compared technique and easy to understand and implement.
- Proposed hybrid technique used simple structure, control flow is well defined and looping structure is also minimized. Due to the following facts it take very less time for execution.

IV. Conclusion

As information systems have become more comprehensive and a higher value asset of organizations, intrusion detection systems has been incorporated as elements of operating systems, although not typically applications. Intrusion detection involves determining that some entity, an intruder, has attempted to gain, or worse, has gained unauthorized access to the system. This research shows that benchmarking intrusion detections systems can be done effectively. In this work design and develop more advanced data mining techniques, it will be very hard to evaluated proposed intrusion detection systems. The amount of customization of data mining techniques that goes into effectively using one, as well as the ever-changing number of viable network exploits makes it impossible at this time. The speed of operation of During Data mining technique is faster. During testing, proposed IDS run in few second to get output. In this no debugging is required. This is due to the high amount of string optimization involved through data mining technique.

References

[1] Om, H. and Kundu, A. "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system" Recent Advances in Information Technology (RAIT), 1st IEEE International Conference on 15-17 March 2012 Page(s):131 - 136 Print ISBN:978-1-4577-0694-3.

[2] P.R Subramanian and J.W. Robinson "Alert over the attacks of data packet and detect the intruders" Computing, Electronics and Electrical Technologies (ICCEET), IEEE International Conference on 21-22 March 2012 Page(s):1028 - 1031 Print ISBN:978-1-4673-0211-1

[3] V. S. Ananthanarayana and V. Pathak "A novel Multi-Threaded K-Means clustering approach for intrusion detection" Software Engineering and Service Science (ICSESS), IEEE 3rd International Conference on 22-24 June 2012 Page(s): 757 - 760 Print ISBN: 978-1-4673-2007-8

[4] N.S Chandolikar and V.D.Nandavadekar, "Efficient algorithm for intrusion attack classification by analyzing KDD Cup 99" Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on 20-22 Sept. 2012 Page(s):1 - 5 ISSN :2151-7681

[5] Virendra Barot and Durga Toshniwal "A New Data Mining Based Hybrid Network Intrusion Detection Model" IEEE 2012.

[6] Wang Pu and Wang Jun-qing "Intrusion Detection System with the Data Mining Technologies" IEEE 2011.

[7] 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 (CITA) 2011.

[8] Dewan M.D. Ferid, Nouria Harbi, "Combining Naïve Bayes and Decision Tree for Adaptive Intrusion detection" International Journal of Network Security and application(IJNSA),vol 2, pp. 189-196, April 2010.

[9] Joseph Derrick,Richard W. Tibbs, Larry Lee Reynolds "Investigating new approaches to data collection,management and analysis for network intrusion detection". In Proceeding of the 45th annual southesast regional conference, 2007. DOI = http://dl.acm.org/citation.cfm?doid=1233341.1233392

[10] M.Panda, M. Patra, "Ensemble rule based classifiers for detecting network intrusion detection", in Int. Conference on Advances in Recent Technology in Communication and Computing, pp 19- 22,2009.

[11] Skorupka, C., J. Tivel, L. Talbot, D. Debarr, W. Hill, E. Bloedorn, and A. Christiansen 2001. "Surf the Flood: Reducing High-Volume Intrusion Detection Data by

All rights reserved by *www.ijsrd.com*

Automated Record Aggregation," Proceedings of the SANS 2001 Technical Conference, Baltimore, MD.
[12] KDD. (1999). Available at http://kdd.ics.uci. edu/databases/ - kddcup99/kddcup99.html
[13] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and regression tres. Monterey, CA: Wadsworth & Books/Cole Advanced Boks & Software, 1984.
[14] Dewan M.D. Ferid, Nouria Harbi, "Combining Naïve Bayes and Decision Tree for Adaptive Intrusion detection" International Journal of Network Security and application(IJNSA),vol 2, pp.189-196, April 2010.
[15] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko Ruth Silverman, Angela Y. Wu " A Local Search Approximation Algorithm for k-Means Clustering" July 14, 2003 Annual ACM Symposium on Computational Geometry.
[16] Eric Bloedorn, Alan D. Christiansen, William Hill "Data Mining for Network Intrusion Detection: How to Get Started" 2001.
[17] Skorupka, C., J. Tivel, L. Talbot, D. Debarr, W. Hill, E. Bloedorn, and A. Christiansen 2001. "Surf the Flood: Reducing High-Volume Intrusion Detection Data by Automated Record Aggregation," Proceedings of the SANS 2001 Technical Conference, Baltimore, MD
[18] Sumathi, S.; Sivanandam, S. N.: Introduction to Data Mining and its Applications. Springer, 2006.
[19] Fayyad, Piatetsky-Shapiro, Smyth: From Data Mining to Knowledge Discovery in Databases. AI MAgazine, 1996.
[20] Roiger, Richard J.; Geatz, Michael W.: Data Mining: A Tutorial- Based Primer. Addison Wesley, 2003
[21] MIT linconin labs, 1999 ACM Conference on Knowledge Discovery and Data Mining (KDD) Cup dataset, http://www.acm.org/sigs/sigkdd/kddcup/index.php?section=1999
[22] The KDD Archive. KDD99 cup dataset, 1999. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

All rights reserved by *www.ijsrd.com*