An Accurate IDS design using KDD CUP 99's Dataset

Ashok Panwar
Technical Officer in ECIL (NPCIL)
Tarapur, Mumbai, India

D. Srinivasa Rao
Associate Prof. in CSE
MITM, Indore, India

G. Sriram
Associate Prof., Dept of CSE
Andhra University,
Visakhapatnam, India

ABSTRACT
IDS or intrusion detection systems are well known network anomaly detection technique in network technology. According to the IDS, it is used for monitoring and analysis of network traffic. By analyzing the network traffic data it observe the behavior of network and report if any anomaly in network behavior occurred. In addition of this technology is also helpful for discovering any attack condition in network. Therefore the proposed work is intended to design and develop an accurate analysis method, which works on KDD CUP 99's Data. The proposed work first involve the feature selection technique using the correlation coefficient based technique and then the selected features are used for training and testing of three popular classifiers namely bays classifier, C4.5 decision tree and KNN algorithm. The experiments are performed using the k-fold cross validation technique. The experimental results shows the KNN and C4.5 decision tree algorithm produces similar accuracy and higher as compared to bays classifier. But the time consumption of the KNN classifier is 10 times higher than the C4.5 and Bays classification techniques.

Keywords
KDD CUP dataset, Classification, data mining, network security, IDS design

1. INTRODUCTION
The network carries very confidential and sensitive kids of data. Attackers are continuously trying to break the security of the network. In this context different security approaches are applied for network security i.e. firewall, anti-viruses, and IDS. The IDS is also known as intrusion detection system. The IDS (intrusion detection systems) are helpful for monitoring the abnormal behavior of network. Therefore the intelligent IDS systems, evaluate the different network parameters and using this probability of network is predicted. This system is usages the data mining and machine learning algorithms, which helps to analyze the historical data and discover the similar patterns from raw data. In this context supervised learning techniques are utilized and using these techniques. The training performed for learning with the attack pattern data and when network observes the similar data pattern the alarm for the critical conditions.

In this presented work a data mining model is presented that provides a way to classify the data using the less amount resource consumption. Therefore to do this task the feature selection approach provided which helps to regulate the number of essential feature selection using the attribute ranking technique. To rank the attributes of data set the correlation coefficient of attributes are measured with respect to the class labels. After measuring the rank the data classification algorithm is applied for classifying the data. To classify and test the algorithm the WEKA library is used for implementing our program with the JAVA. The KNN (k-nearest neighbor), bays classifier and C4.5 decision tree is applied for experimentation.

This section provides the overview of the proposed intelligent IDS (intrusion detection system) design. The next section provides the objective of conducted experiments, data model for classifying patterns. In addition of that the experimental results are provided. Finally the paper provides the conclusion and final future extension of the proposed work.

2. PROPOSED WORK
This section provides the detailed discussion on the proposed IDS (intrusion detection system) design. In this context this section includes the overview of functional system, methodology and the algorithm by which the classification task is performed.

2.1 System overview
The data mining and its techniques are one of the powerful tools of this computational era. New contributions and applications are developed for various domains. These techniques are used for prediction, finding similar patterns (those values are not exactly same as required but following a target pattern), clustering and others. Therefore, a number of application areas accepting this technology. In this presented paper an application of data mining in the domain of security and network behavior analysis is presented. The network services require a system known as IDS (intrusion detection system) utilizes. That application is used for security purpose. Those applications analyze the network behavior parameters and label the data among malicious and legitimate patterns.

In this presented work IDS (intrusion system design) using the data mining techniques is presented. That is focused on the following objectives:

1. Finding the accurate algorithm for which is best fit for classifying network behaviour patterns
2. Finding the additional parameters by observation of experimental results on the following basis:

a. Which algorithm is accurate for classification

b. Which algorithm consumes more time for classification of target patterns

c. Which algorithm is higher memory resource consuming

d. Which algorithm is best for all the parameters

In this context a data model in next section is explained which helps to accept the experimental data, reducing the noise from the data, reducing dimensions of the data and classifying the data accurate manner. The required algorithms their utilization sequence is presented in the next section.

2.2 Methodology

The proposed methodology of the system design is defined using the figure 1. The required components for processing the data are also demonstrated in this diagram.

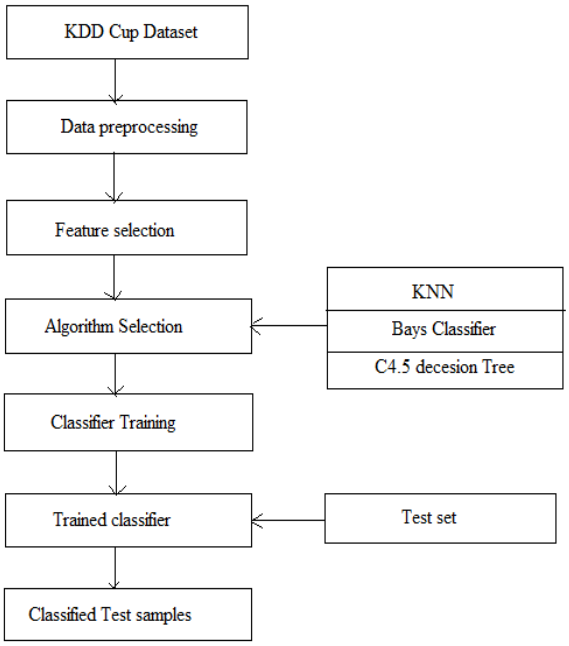

Fig. 1 proposed system architecture

KDD cup dataset: the KDD cup dataset is basically a network trace file which contains a significant amount of data instances. Additionally it contains 41 attributes and one class label. The entire dataset is sub-divided in four major classes. These attributes are based on the network parameters and the class labels represent the different kinds of attacks which are based on network behavior. Therefore this dataset is frequently used for designing and development of IDS (intrusion detection systems). In this presented work total 15000 data instances from all classes of attack are selected for system design and experimentation.

Data preprocessing: the data preprocessing techniques are used for improving the quality of dataset instances. The aim of preprocessing is also impact on the classifiers performance. Therefore different techniques can also be applied with the dataset for this purpose. In this context the data cleaning, data transformation, removal of null values are essential process. In this presented work the data preprocessing technique is applied for identifying the missing attributes of the dataset. Additionally these data instances which contain the missing values are removed by the system.

Feature selection: as described in description of KDD cup dataset, the dataset contains a huge amount of information therefore the dimensions of the dataset is too much higher. In this context for identifying the essential features among the available data the feature selection technique is applied. To identify the much valuable features from the data correlation coefficient technique is applied to rank the attributes. After ranking of attributes the essential target features are preserved and remaining data is removed from dataset. Using this approach the computational resources are also preserved in terms of time and memory resources. In order to rank the attributes the following formula is used by which relevancy of attributes with respect to class labels is measured.

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Where n is the total number of data instances in dataset

x_i, y_i are the attributes and class values with index i

$\bar{x}, \bar{y}$ is the mean of attribute values

Algorithm selection: the three most popular algorithms are selected for implementation and performing the classification. The aim of the implementation of this algorithm is to find most fit classifier for designing the IDS (intrusion detection system). The description of the implemented algorithms is give as follows:

KNN: the KNN algorithm is also known as the k-nearest neighbor classifier. This classification algorithm is working on the basis of distance based method by comparing the target instances with the entire dataset instances. There are the following distance formula is used for processing the test data with respect to the training data.

$$d(x, y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

Where, the x and y the query and database sequences for comparing each other.

Bays classifier: the bays classifier is probability based classification approach. That computes the posterior and priori probability to estimate the classes of the data instances produced for classification. There are two types of probability as follows:

- Posterior Probability [P (H/X)]
- Prior Probability [P (H)]

Where, X is data tuple and H is some hypothesis. According to Baye's Theorem

$$P\left(\frac{H}{X}\right) = \frac{P\left(\frac{X}{H}\right)P(H)}{P(X)}$$

C4.5 decision Tree: the C4.5 decision tree algorithm is a rule based classification algorithm. That algorithm provides the decision tree using the available attribute values. That approach is advance version of the ID3 decision tree algorithm. In this tree the attributes are mounted on tree branches and leaf nodes are used for providing the decisions for the applied test instances.

Classifier Training: all these algorithms are used for processing data and using these algorithms the data models are developed.

Trained classifier: the data models are developed using the applied classifiers and training samples, which is used for classification of the instances. These data models are termed here as the trained classifiers.

Test dataset: the selected features are initially divided into two parts training set and testing set. The training set is prepared using the 70% of randomly selected dataset instances. Additionally the 30% of remaining randomly selected data is used for testing of trained classifier.

Classified test samples: the test dataset is applied on trained data models, which is classified using the algorithms which

are trained. After classification the class labels of the data is identified according to the algorithm processes. Additionally using the classified test samples the performance of the classification algorithms are measured.

2.3 Proposed algorithm

The proposed system model is demonstrated in the previous section. That model is summarized using the algorithm steps which are represented in table 1.

Tables 1 proposed algorithm

Input: KDD CUP dataset D, Selected algorithm SA, Target feature size FS, Test dataset T

Output: class labels identified C

Process:

1. $[attributes, instances] = readDataset(D)$
2. $for(i = 1; i \leq instances; i++)$
   a. $if(D_i.contains(null))$
      i. $remove(D_i)$
   b. $else$
      i. $PData.Add(D_i)$
   c. $end\ if$
3. End for
4. $for(j = 1; j \leq PData.Size; j++)$
   a. $Rank_j = ComputeCorrelation(PData_j, Class_j)$
5. $end\ for$
6. $FeatureF = selecteAttributes(PData, Rank, FS) T_{model} = SA.Train(FeatureF)$
7. $C = T_{model}.Classify(T)$
8. Return C

3. RESULT ANALYSIS

This section provides the complete details about the measured performance parameters for the proposed data mining based IDS system design.

3.1 Accuracy

The accuracy is the measurement of correctness of a data mining algorithm. Therefore that is computed on the basis of total correctly classified samples and total samples to classify. The following formula is used for calculating the accuracy of algorithms.

$$accuracy = \frac{total\ correctly\ classified\ samples}{total\ samples\ to\ classify} X100$$

Tables 2 accuracy

S. No.	KNN	Bays	C4.5
1	91	81	91
2	93	87	96
3	94	83	93
4	96	89	96
5	93	86	95
6	97	83	94
7	99	91	92

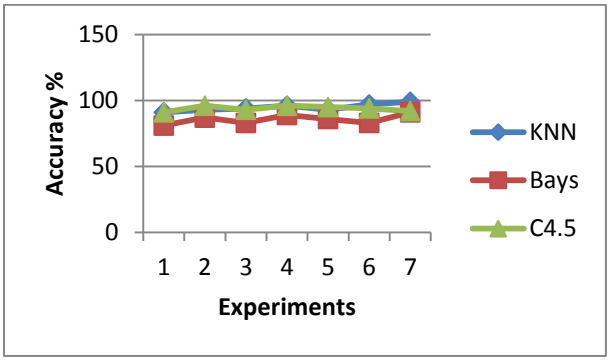

Fig. 2 accuracy

The accuracy of the proposed work is measured for all the implemented algorithms. The comparative performance in terms of accuracy is given using figure 2 and table 2. The measurement of the accuracy is provided here in terms of percentage. In this diagram the X axis shows the experiments conducted with the algorithms and the observed accuracy is given in Y axis. According to the obtained performance the C4.5 algorithm is providing a consistent performance additionally the KNN classifier provides the higher accuracy during observations. Finally the bays classifier is low performing algorithm for proposed scenario of data classification. Thus C4.5 and KNN is recommended for further classification task with KDD CUP Dataset.

3.2 Error Rate

The error rate of the data mining system indicates the misclassified samples among the total samples produced for classification. That is sometimes also called the misclassification rate of classification algorithms. The following formula can be used for calculating the error rate.

$$error\ rate = \frac{total\ misclassified\ samples}{total\ samples\ to\ classify} X100$$

Tables 3 error rate

S. No.	KNN	Bays	C4.5
1	9	19	9
2	7	13	4
3	6	17	7
4	4	11	4
5	7	14	5
6	3	17	6
7	1	9	8

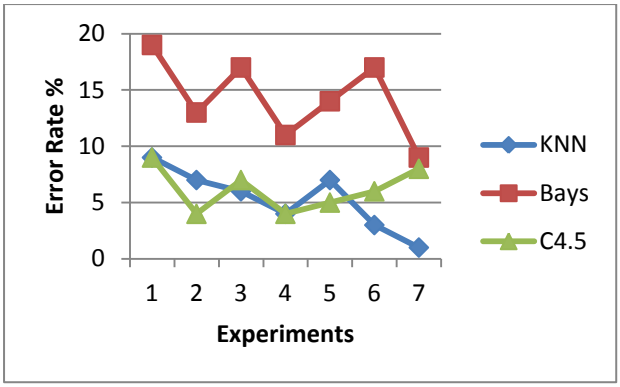

Fig. 3 error rate

The computed error rate of the system for all three implemented algorithms i.e. KNN, Bays and C4.5 decision tree is mentioned in table 3 and figure 3. The calculated error rate of the system is demonstrated in terms of percentage. The X axis of figure 3 shows the experiments performed with the implemented algorithms and the Y axis shows the obtained error rate of algorithms in terms of percentage. According to the obtained performance the KNN and C4.5 shows the less error rate as compared to the Bays classifier. But the KNN produces less error then C4.5 algorithm. But the consistency of the error rate is C4.5 is higher as compared to other two algorithms.

3.3 Memory Usage

The memory usages of the system are the total amount of main memory which is acquired by the algorithm for processing the data. The measurement of memory usages in JAVA technology is performed using the following formula.

$$memory\ usage = assigned\ memory - freememory$$

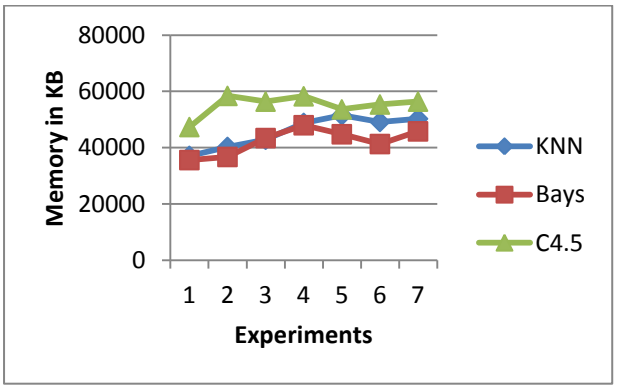

Fig. 4 memory usages

Tables 4 memory usages

S. No.	KNN	Bays	C4.5
1	37114	35618	47346
2	40284	36727	58419
3	42847	43425	56381
4	48821	47991	58277
5	51538	44793	53628
6	49173	41313	55372
7	50288	45811	56382

The memory usages of the given IDS detection system is provided using figure 4 and table 4. The memory usage of the implemented algorithms is measured in terms of KB (kilobytes). The X axis of the represented diagram demonstrates the experimental observations and the Y axis shows the memory usages of the implemented system using three different algorithms. According to observed memory usages of the algorithms the C4.5 decision tree usages higher memory as compared to other two algorithms. That is because the tree needs to store all the training data into the main memory directly therefore the algorithm requires higher amount of memory as compared to other two algorithms.

3.4 Time Consumption

The time consumption of the proposed data model is demonstrated in this section. The time required to classify all the test data samples are measured in terms of time consumption. It is computed using the following formula.

$$time\ consumption = end\ time - start\ time$$

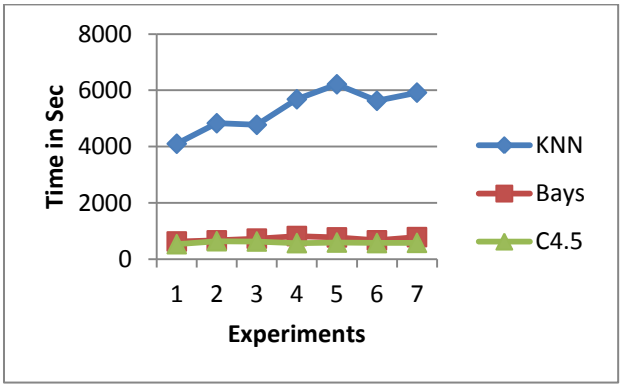

Fig. 5 time consumption

Tables 5 time consumption

Experiments	KNN	Bays	C4.5
1	4101	627	533
2	4830	667	639
3	4773	729	628
4	5682	821	571
5	6214	772	591
6	5627	668	572
7	5918	782	580

The total amount of time consumed for classifying data with the three different implemented algorithms are represented using table 5 and figure 5. In this diagram the X axis shows the experiments performed with the system and Y axis demonstrate the consumed time. The time measured in this system is given here in terms of seconds. According to the calculated time the KNN algorithm is much complicated and expensive. Additionally both the algorithms are consuming very less amount of time as compared to KNN algorithm..

4. CONCLUSION AND FUTURE WORK

This section provides the summary of the proposed IDS system design using the data mining techniques. Thus based on the experimentations the conclusion of the proposed system is defined and the future extension of the proposed work is also discussed in this section.

4.1 Conclusion

The large organizations are always worried about the security of their internal and external network security. Therefore the security and monitoring of these network systems are required. In such conditions the attack may possible on both the ends, by using internal network and also by using the external attacker. Therefore the sampling and evaluation of the different network patterns is essential for large and secure networks. In this context the IDS system can be helpful for monitoring and discovering the anomaly in network by using the network parameters and their fluctuation analysis in time to time basis.

Therefore in this presented work an IDS system is proposed for design and implementation. The given design of the IDS system is easy to understand, efficient and accurate for network behavior classification. The given model works in two phases' first training of the system and then testing. The training and testing of the given data model is performed on 15000 samples of KDD cup dataset instances. Additionally to reduce the resource consumption in terms of time and memory the feature selection technique is used. After selection of effective parameters the classifiers are applied and the performance of system is measured.

The implementation of the proposed technique is performed using JAVA technology. Additionally to implement the classification algorithms the WEKA library is used. After implementation of the proposed working system the performance of the system is measured and the summary is reported using table 6.

Tables 6 performance summary

S. No.	Parameters	KNN	Bays	C4.5
1	Accuracy	91-99 %	81-91 %	91-96 %
2	Error rate	1-9 %	9-19 %	4-6 %
3	Time consumption	4101-6214 S	627-821 S	533-639 S
4	Memory usage	37114-51538 KB	35618-47991 KB	47346-58419 KB

According to the obtained performance the proposed system found acceptable and accurate for classifying the network behavioral attributes. Therefore the proposed technique can be acceptable for real world IDS system design and development.

4.2 Future Work

The proposed model is implemented and tested successfully based on the performance obtained it is accurate for classification of network behavioral data. Therefore the following work is proposed for future implementation and design.

1. Evaluation of new generation techniques of machine learning i.e. deep learning, convolution neural network
2. Reduce the dimensions more to preserve the time and memory

5. REFERENCES

[1] K. Kendall, A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems, PhD thesis, MIT Lincoln Laboratory, 1999.

[2] S. Archana and Dr. K. Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Volume 2 Issue 2, February 2014.

[3] Han, Jiawei, Jian Pei, and Micheline Kamber, Data mining: concepts and techniques, Elsevier, 2011.

[4] Saranya Vani. M and Dr. S. Uma, "Survey on Classification Techniques Used in Data Mining and their Recent Advancements", International Journal of Science, Engineering and Technology Research, Volume 3, Issue 9, September 2014

[5] Senthilnayaki Balakrishnan and Venkatalakshmi K, "Intrusion Detection System Using Feature Selection and Classification Technique", International Journal of Computer Science and Application (IJCSA) Volume 3 Issue 4, November 2014

[6] Snehal A. Mulay and P.R. Devale, "Intrusion Detection System using Support Vector Machine and Decision Tree", International Journal of Computer Applications (IJCA), Volume 3 – No.3, June 2010

[7] Zibusiso Dewa and Leandros A. Maglaras, "Data Mining and Intrusion Detection Systems", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, 2016

[8] Mukkamala, Srinivas, Guadalupe Janoski, and Andrew Sung, "Intrusion detection using neural networks and support vector machines", Proceedings of the International Joint Conference on Neural Networks, IJCNN'02, Volume 2, IEEE, 2002

6. AUTHORS PROFILE

ASHOK PANWAR received has Three Year Polytechnic Diploma in Computer Science and Engineering, B.E. / B. Tech. and M.E. / M. Tech. Degree both in Computer Science and Engineering. He is Currently working as an Technical Officer in **ECIL** (Electronics Corporation of India Limited), Hyderabad, India, against the site requirements of **NPCIL** (Nuclear Power Corporation of India Limited), Tarapur, Mumbai, Maharashtra, Working in **ACS** (Access Control System) Department, as well as Research Scholar, Ex. Employee in Defence Research & Development Organisation **(DRDO)** in Defence Scientific Information & Documentation Centre **(DESIDOC)** Lab, Govt. of India, Ministry of Defence, in Department of Knowledge Management Division **(KMD)**, Metcalfe House, Near Civil Lines, New Delhi, Delhi-110054, India. He has one year of Teaching Experience in Computer

Networking. His area of Main Research Interest in Ad-hoc Networks, Network Attacks, MANET, Data Mining & Network Security. He has guided 15 Graduate Students. He has published 01 paper in international journal. He has attended One National Level Conference. He has attended Two National Level Event's of Microsoft Dream Spark Yatra at IET - DAVV, Indore. He has attended Three Day's CEP on **Information Security in Web Based Services** organised by **DRDO (DESIDOC)**, New Delhi. He has attended Workshop on **DRDO E-journals Service** organised by **DRDO (DESIDOC)**, New Delhi. He has attended Five Day's National Level Workshop on Android Security System. He has attended Two Day's National Level Workshop in **NS2 (**Network Simulator and Design 2) at MITM, Indore and attended Three Day's National Level Seminar.

D. SRINIVASA RAO M.Tech, Ph.D is working as an Associate Professor in the Department of Computer Science & Engineering at Medi-Caps University, Indore, Madhya Pradesh, India. He has 22 years of teaching experience. His area of interest in Adhoc Networks, Distributed Systems, Network Security & Image Processing. He has guided more than 60 Post Graduate Students. He has published 2 books and 18 papers in international journals. He presented 2 papers in National Conferences, 1 paper in International Conference and has attended 37 National Workshops / FDP / Seminars etc. He is a life member of Professional Society like ISTE.

G. SRIRAM M.Tech, Ph.D is working as an Assistant Professor in the Department of Computer Science, School of Distance Education, Andhra University, Visakhapatnam, India. He has 13 years of teaching experience. His area of interest in Adhoc Networks, Data Mining & Networks Security. He has guided 25 Graduate Students. He has published 5 papers in international journals. He has attended 10 National Workshops / FDP / Seminars etc.