Improved Performance of Network Attack Detection using Combination Data Mining Techniques

Amrit Priyadarshi
Research Scholar, CSJMU, Kanpur, India

Dr. Rashi Agarwal
Reader, CSJMU, Kanpur, India

***Abstract* --** Network Attack detection is very important mechanism for detecting attack in computer networks. Data mining techniques play very important role in detecting intrusions in computer networks. Intrusions can damage to the data and compromise integrity and confidentiality and availability of the data. Intrusions are the activities that violate the security policy of system. Intrusion Detection is the process used to identify network attack. Network security is to be considered as a major issue in recent years, since the computer network keeps on expanding every day. A Network Attack Detection System (NADS) is a system for detecting intrusions and reporting to the authority or to the network administration. Data mining techniques have been applied in many fields like Network Management, Education, Science, Business, Manufacturing, Process control, and Fraud Detection. Data mining algorithms like J48, Randam Forest ,Random Tree, Hoefding Tree and Rep Tree are used to build intrusion detection models using KDD CUP 1999. The performance of network attack detection model is evaluated using KDD CUP 1999 test dataset using series of experiments and measured using correct classification and detection of attack. The combination of data mining algorithm will increase performance of network attack detection i.e false positive and false negative, novel or unknown attacks.

***Keywords* –** *NADS; Data Mining; Intrusion detection; IDS; J48; Random Forest; Random Tree; Hoefding Tree; Rep Tree;*

*****

I. INTRODUCTION

An network attack is defined as type of action which compromises the integrity, confidentiality or Availability. Although it plays a very important role to define and protect in security architecture, but NADS is still immature and not considered as a complete defense,. NADS identifies or monitors any kind of attack and notify immediately in the form of alert so that resources never get compromised. An NADS is also used in legal proceedings as forensic evidence against the intruder because it provides recording of any kind of intrusion involved in cybercrime. An NADS is deployed to cover unauthorized access to resources or data. It can be hardware and/or software. An NADS can be used to protect a single host or a whole computer network. NADS which provides user friendly interface to non-expert staff for managing the systems easily. Network attack is any kind of unauthorized activity on a computer network .It is achieved passively or actively. In passive, intrusion takes place by information gathering whereas in case of active intrusion takes place through harmful packet forwarding, packet dropping and by hole attacks [1]. An NADS is a process or device that monitors events occurring on a network and analyzing it to detect any kind of activity that violate computer security policies. The NADS device can be hardware, software or a combination of both that monitors the computer network against any unauthorized access [2]. The main motive of the NADS is to catch the intruder before a real and serious damage to computer network.

II. DATA MINING

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers “to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases” [3]. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure 1 shows data mining as a step in an iterative knowledge discovery process [4].

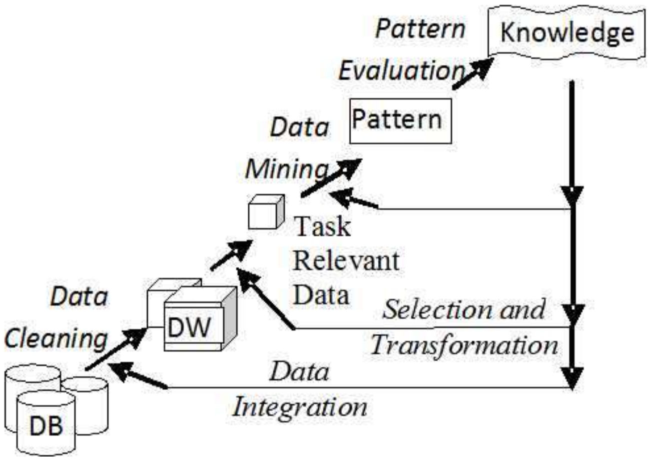

Figure 1. Data Mining the core of Knowledge Discovery process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps[5]:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: It is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data. The KDD is an iterative process [6]. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results. Classification techniques are based on establishing an explicit or implicit model that enables categorization of network traffic patterns into several classes [7][8]. Analysis of the KDD dataset showed that there were two important issues in the dataset, which highly affect the performance of evaluated systems resulting in poor eval- uation of anomaly detection methods [9]. To solve these issues, a new dataset known as NSL-KDD [10], consisting of selected records of the complete KDD dataset was introduced. This dataset is publicly available for researchers on http://www.iscx.ca/NSL- KDD/ and has the following advantages over the original KDD dataset.

III. NETWORK ATTACK DETECTION SYSTEM

Figure 2 represent the high-level system architecture. The system will be constructed from multiple distinct components:

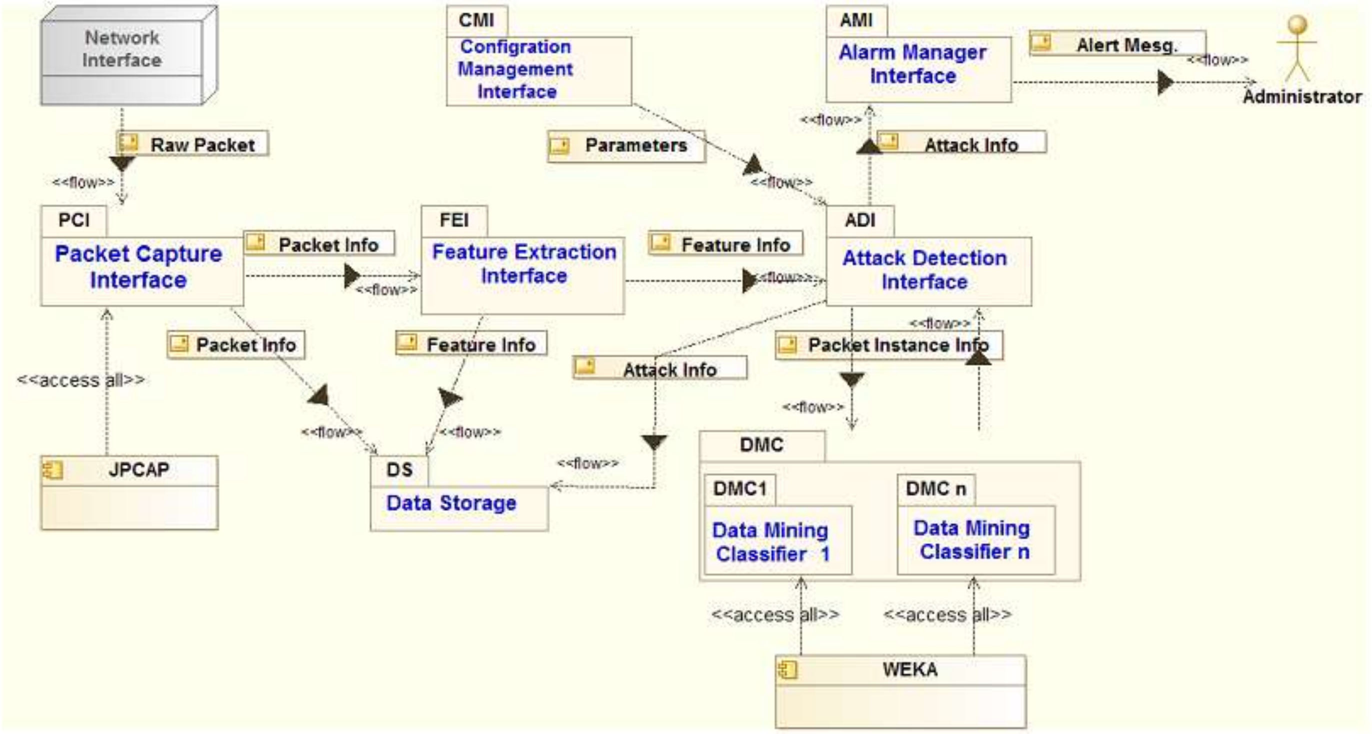

Figure 2. System Architecture

Network Interface:

A network Device/Sensor is a network interface with a communications infrastructure intended to transmit and receive network traffic.

Attack Detection Engine :

It consist of four components : Packet Capture Manager, Feature Extraction, Data Mining Classifiers and Anomaly Detector.

Packet Capture Interface :

It provides interface for processing of raw packets receives at network sensors. It analyzes TCP and UDP packet and translate them into the required form. It is dependent on JPCAP library. It also provide interface for storing packet information into the database.

Feature Extraction Interface :

It provide interface for extracting features such as duration, protocol, service etc from captured packet. It also provide interface for storing feature information into the database.

Attack Detection Interface :

It provides interface for classifying packet instance using Data Mining Classifier Interface and system configuration.

Data Mining Classifier (s) Interface :

It provides interface to Data Mining classification models and provides methods for classification of packet instance. It depends upon WEKA library.

Alert Manager interface :

Send or Display alert messages.

Configuration Management Interface:

Provide interface for configuring system parameters such as protocol for packet capture, data mining algorithm for attack detection and Alert messages and stores them.

Data Storage :

Stores packet information, attack information and alarm messages in to database.

JPACP :

Java packet capture library.

WEKA :

Collection Data mining algorithms in Java.

IV. RESULT AND DISCUSSION

The result shown in Table 1. shows overall classification accuracy in terms of correctly classified and wrongly classified record of test data set. Experiment is performed single Data Mining Technique which is shown Table 2. and we performed another experiment by combining two classifiers and results obtained are show in Table 3. and Table 4. These table shows percentage of correct attack and false attack detection for different combination of algorithms. It is clear that the combination of Random Forest with Hoeffding Tree combination performs comparatively better than any other combination in PROBE and Normal attack category. Similarly Random Tree with Rep Tree, Random Tree with Hoeffding Tree and Hoeffding Tree with REP Tree combination performs comparatively better than any other combination in DOS, R2L, U2R attack category respectively.

Overall this combination performance was improved more than 1% in correct attack detection compared to single best algorithm.

Table: 1. Performance matrices

Classifiers	Classified Instances	
	Correctly	Incorrectly
J48	74.7028	25.2972
Random Forest (RF)	77.8921	22.1079
Random Tree (RT)	74.2814	25.7186
Hoeffding Tree (HT)	79.0454	20.9546
REPTree (RepT)	75.3504	24.6496

Table 2. Percentage of Attack Detection using Sngle Data Mining Algorithm

CLASSIFIER	ATTACK TYPES					
	DOS	PROBE	R2L	U2R	NORNAL	OVERALL
J48	76.026	64.519	6.235	13.433	97.003	74.703
RANDOM FOREST(RF)	82.153	73.276	7.101	4.478	97.323	77.892
RANDOM TREE(RT)	76.629	66.956	10.010	25.373	93.749	74.281
HOEFDING TREE(HT)	81.315	78.645	26.290	34.328	93.379	79.045
REP TREE(RT)	82.220	69.021	10.703	47.761	91.062	75.350

We conclude that combination of Hoeffding tree and REP Tree performs better than other combination. The performance of the system was improved because Hoeffding tree performs better in PROBE and R2L type of attack and J48 and Random Forest performs better in Normal type and REP Tree performs better in U2R Category.

Thus by combining advantages of each classifier we can achieve better attack detection rate and able to reduce false attack detection rate. Despite the improvements in most of the category none of combination has achieved improvements in all the category.

Table 3. Percentage of Correct Attack Detection using Combination of Data Mining Algorithm

Classifiers	Attack Types				
	DOS	R2R	PROBE	U2R	Normal
J48 & RF	82.743	8.929	76.749	17.910	97.364
J48 & RT	79.056	12.539	74.308	31.314	97.436
J48 & HF	82.703	26.602	85.998	41.791	97.384
J48 & RepT	83.052	11.742	75.836	50.746	97.158
RF &RT	84.245	12.019	76.332	26.866	97.539
RF &HT	82.757	26.429	88.765	34.326	97.559
RF &RepT	82.931	10.876	77.034	49.254	97.354
RT & HT	84.460	27.884	85.171	43.284	97.057
RT &RepT	84.621	13.929	70.136	55.224	93.842
HT &RepT	82.877	26.741	85.832	62.687	93.945

Table 4. Percentage of False Attack Detection using Combination of Data Mining Algorithm

Classifiers	Attack Types				
	DOS	R2R	PROBE	U2R	Normal
J48 & RF	17.257	91.271	23.255	82.090	2.636
J48 & RT	20.944	87.461	25.692	68.657	2.564
J48 & HF	17.297	73.398	14.002	58.209	2.616
J48 & RepT	16.948	88.258	24.164	49.254	2.842
RF &RT	15.755	87.981	23.668	73.134	2.461
RF &HT	17.243	73.571	11.235	65.672	2.441
RF &RepT	17.069	89.124	22.966	50.746	2.646
RT & HT	15.540	72.116	14.829	56.716	4.943
RT &RepT	15.379	86.076	29.864	44.776	6.158
HT &RepT	17.123	73.259	14.168	37.313	6.055

Figure 3 and Figure 4 shows graphical representation of percentage of correct attack detection and percentage of false attack detection using combination of algorithms respectively. Most of the combination achieve better performance than single algorithm because single algorithms can‘t perform better in all types of attack. None of combination performs better in R2L and U2R category because number of records in training set are very less compared to test data set.

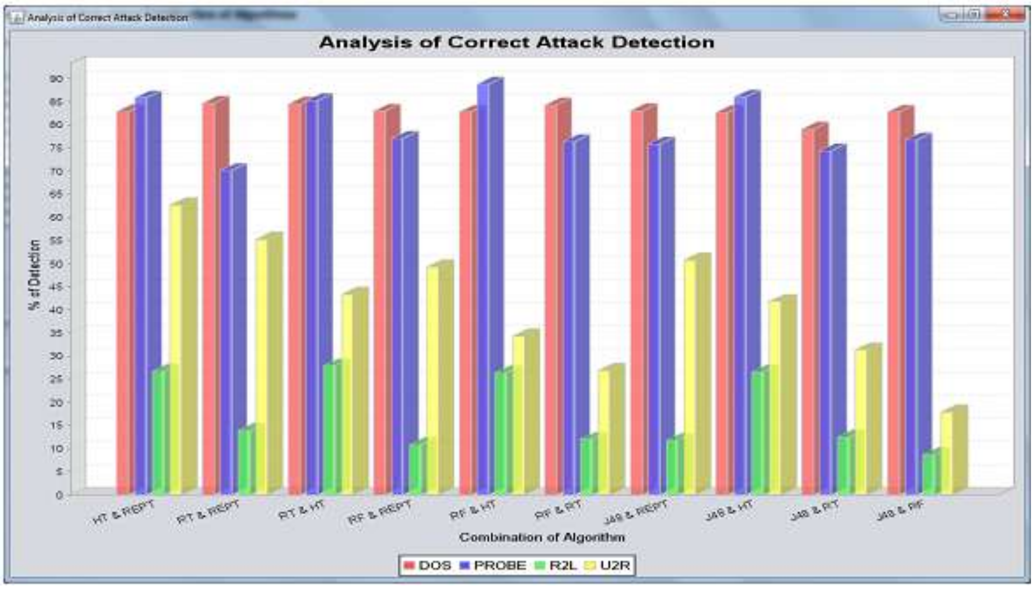

Figure 3. % of correct attack detection using combination of data mining algorithms

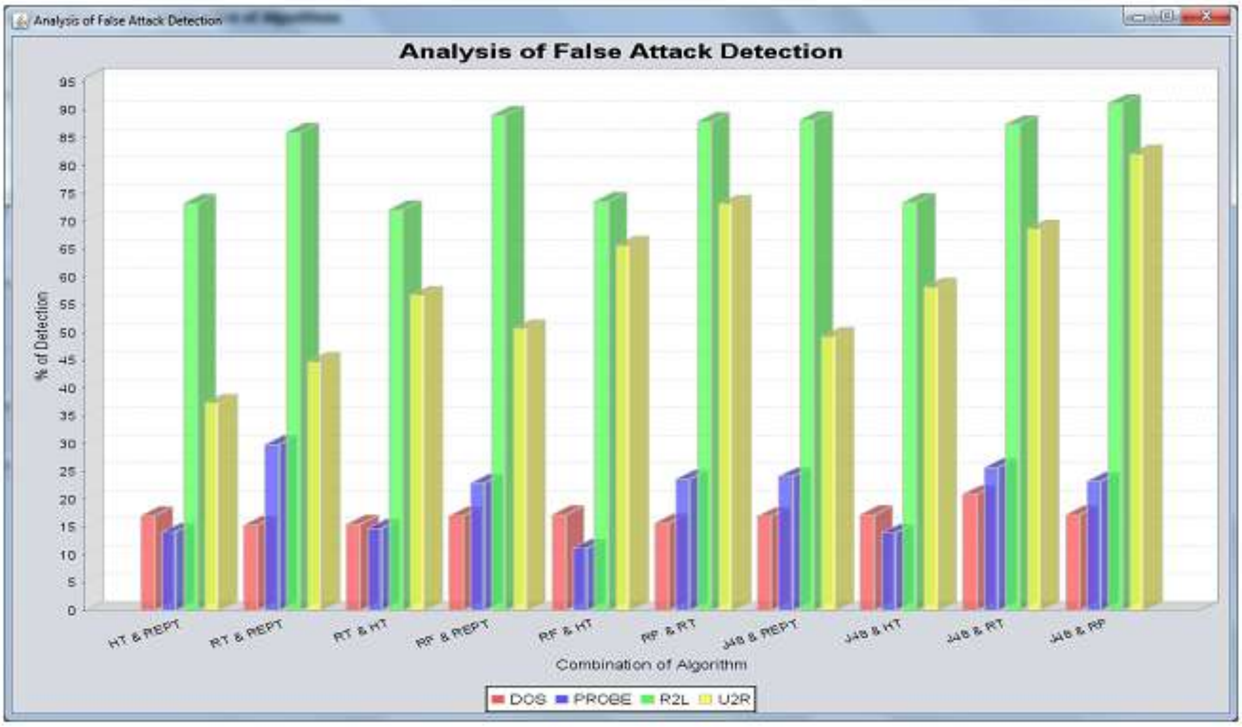

Figure 4. % of false attack detection using combination of data mining algorithm

V. CONCLUSION

This study has found that generally combination of two or more data mining techniques will improve performance of attack detection ratio significantly compared to single data mining technique. We have presented evaluation results by combining J48, Random Tree, Random Forest, Hoeffding Tree and REP Tree with each other using NSL-KDD dataset.

REFERENCES

[1] H. Debar, M. Dacier, and A. Wespi, "A revised taxonomy for intrusion-detection systems," In Annales des telecommunications, Springer-Verlag, vol. 55, no. 7-8, pp. 361-378, 2000.

[2] A.S. Ashoor and S. Gore, "Importance of Intrusion Detection system (IDS)," International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp. 1-4, 2011.

[3] W. J. Frawley, G. P. Shapiro, and C. J. Matheus, “Knowledge discovery in databases: An overview,” AI Magazine, vol. 13, pp. 213–228, 1992.

[4] P. Sundari and D. K. Thangadurai, “An empirical study on data mining applications,” Global Journal of Computer Science and Technology, vol. 10, no. 5, 2010.

[5] O. R. Zaian, Introduction to Data Mining. University of Alberta,Canada, 1999. [30]W. Lee and S. J. Stolfo, “A framework for constructing features and models for intrusion detection systems,” ACM Transactions on Information and System Security, vol. 3, pp. 227–261, 2000.

[6] KDDcup99, Knowledge discovery in databases DARPA archive, 1999.

[7] F. S. Wattenberg, J. I. A. P. rez, P. C. de la Higuera, M. M. M. n Ferna ndez, and I. A. Dimitriadis, “Anomaly detection in network traffic based on statistical inference and a stable modeling,” IEEE Trans Dependable Secure Computing, vol. 8, pp. 493–509, July/August 2011.

[8] H. Tong, C. Li, J. He1, J. Chen, Q. A. Tran, H. Duan, and X. Li, “Anomaly internet network traffic detection by kernel principle component classifier,” vol. LNCS, pp. 476–481, 2005.

[9] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A detailed analysis of the kdd cup 99 data set,” in Proc 2nd IEEE International Conference on Computational Intelligence for Security and Defense Applications USA IEEE Press , pp. 53–58, 2009.

[10] N. KDD, NSL KDD data set for network-based intrusion detection systems, March 2009.