ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2, No 5, May 2013

Important Roles Of Data Mining Techniques For Anomaly Intrusion Detection System

Phyu Thi Htun and Kyaw Thet Khaing

Abstract*— Today, there are so many information interchanges are performed in that internet as the increasing the amount of using internet. That why, the methods used as intrusion detective tools for protecting network systems against diverse attacks are became too important. The available of IDS are getting more powerful. But, modern intrusion detection system facing complex problems. These system has to be require reliable, extensible, easy to manage, and have low maintenance cost. In recent years, data mining-based intrusion detection systems (IDSs) have demonstrated high accuracy, good generalization to novel types of intrusion, and robust behavior in a changing environment. In this proposed model, we focus on the best two data mining algorithms for intrusion detection system. The k-Nearest Neighbor was used as the classical pattern reorganization tools have been widely used for Intruder detections. There have some different characteristic of features in building an Intrusion Detection System. Conventional k-NN do not concern about that. Our enhanced Model proposed with an Random Forest (RDF) and k-Nearest Neighbor (kNN) method to perform more accurate classification task of the new model. RDF can select more important features and kNN can select more precisely than the conventional System. Experiments and comparisons are conducted through intrusion dataset: the KDD Cup 1999 dataset.

***Index Terms*— Intrusion Detection System, Random Forest,and k-Nearest Neighbor KDD Cup. unknown(novel)attack**

I. Introduction

This Intrusion Detection Systems (IDSs) have become a major focus of computer scientists and practitioners as computer attacks have become an increasing threat to commercial business as well as our daily lives. Researchers have developed two main approaches for intrusion detection: misuse and anomaly intrusion detection. Misuse consists of representing the specific patterns of intrusions that exploit known system vulnerabilities or violate system security policies.

Manuscript received May, 2013.

***Phyu Thi Htun**, Faculty of Information and communication Technology, University of Technology ,Yatanarpon Cyber City, Myanmar., Yatanarpon Cyber City, Myanmar, Phone/ Mobile No: +959402560402*

***Kyaw Thet Khaing**, Computer Hardware Department, University of Computer Studies, Yangon,Yangon, Myanmar,*

On the other side, anomaly detection assumes that all intrusive activities are necessarily anomalous. This means that if we could establish a normal activity profile for a system, we could, in theory, flag all system states varying from the established profile as intrusion attempts. These two kinds of systems have their own strengths and weaknesses.

The former can detect known attacks with a very high accuracy via pattern matching on known signatures, but cannot detect novel attacks because their signatures are not yet available for pattern matching. The latter can detect novel attacks but in general for most such existing systems, have a high false alarm rate because it is difficult to generate practical normal behavior profiles for protected systems.

In this paper, we only consider anomaly detection systems, extend the definition of anomaly detection to not only take into account normal profiles but also handle known attacks and explore supervised machine learning techniques. These techniques have proven their efficiency in predicting the different classes of the unlabeled data in the test data set for the KDD'99 intrusion detection contest.

The rest of the paper is organized as follows. Section 2 presents the related works using corresponding machine learning Algorithms. Section 3 described the KDD 99 intrusion detection cup dataset. Section 4 introduces machine-learning techniques, Random Forests and k-Nearest Neighbor, and presented the proposed system model. Using those machine learning algorithms in our proposed system, which presented in Section 4, Section 5 describes the experimental results obtained by using the machine-learning algorithms with WEKA tool[15]. This results obtained with the algorithm over KDD99 do not correspond to what we expect. This is due, in reality, to the transformation of DARPA 98 to KDD 99. Section 6 concluded for our research by using the out coming results using with those machine learning algorithms

II. Procedure for Paper Submission

A IDDM (Intrusion Detection using Data Mining Techniques) [24] is a real-time NIDS for misuse and anomaly detection. It applied association rules, Meta rules, and characteristic rules. Jiong Zhang and Mohammad Zulkernine [21] employ random forests for intrusion detection system. Random forests algorithm is more accurate and efficient on large dataset like network traffic. We also use this data mining technique to select features and handle imbalanced intrusion problem. The most related work to ours is done also by them [19]. They use Random Forests Algorithm over rule-based NIDSs. Thus, novel attacks can be detected in this network intrusion detection system.

In contrast to the previously proposed data mining based IDSs, we employ random forests for anomaly intrusion detection. Random forests algorithm is more accurate and efficient on large dataset like network traffic. We also use the data mining techniques to select features and handle imbalanced intrusion problem.[16]

Random Forest (RDF) also intend to handle new instances that are not considered in all current supervised machine learning techniques[21], And k- Nearest Neighbor(k-NN) algorithm, is one of those algorithms that are very simple to understand but works incredibly well in practice. k-NN method was used as a supporter method for multi-class classification [22][25].

III. DATASET DESCRIPTIONS

A Since 1999, KDD'99 [12] has been the most widely used data set for the evaluation of anomaly detection methods. This data set is built based on the data captured in DARPA'98 IDS evaluation program [8]. DARPA'98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic. The two weeks of test data have around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type. The simulated attacks fall in one of the following four categories:

(1) Denial of Service Attack (DoS): is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.

(2) User to Root Attack (U2R): is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.

(3) Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.

(4) Probing Attack: is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls. Table 1 showed the four categories and their corresponding attacks on each categories.

Table I. Classification of attacks on KDD dataset

Classification of Attacks	**Attack Name**
DoS	smurf, land, pod, teardrop,
R2L	ftp_write, guess_passwd, imap,
U2R	perl, buffer_overflow, rootkit,
Probe	ipsweep, nmap, satan,

It is important to note that the test data is not from the same probability distribution as the training data, and it includes specific attack types not in the training data which make the task more realistic. Some intrusion experts believe that most novel attacks are variants of known attacks and the signature of known attacks can be sufficient to catch novel variants.

The KDD CUP shared 4 dataset file, Train+, Train+_20Percent,Test+ and Test-21. The first two files represent for training datasets and contain the general attacks. The rest two files represent for testing datasets and contain not only general attacks but also the unknown (novel) attacks. The connection for each attack type is shown in Table II.

Table II. Number of connection in each attack type

Datasets	Normal	DoS	U2R	R2L	Probe	Total
Train+	67343	45927	993	54	11656	125973
Train+20 Percent	13449	9234	206	12	2289	25190
Test+	9711	7458	2421	533	2421	22544
Test-21	2152	4342	2421	533	2402	11850

IV. MACHINE LEARNING ALGORITHMS

To overcome the limitations of the rule-based systems, a number of IDSs employ data mining techniques. Data mining is the analysis of (often large) observational data sets to find patterns or models that are both understandable and useful to the data owner [17][23]. Data mining can efficiently extract patterns of intrusions for misuse detection, establish profiles of normal network activities for anomaly detection, and build classifiers to detect attacks, especially for the vast amount of audit data. Data mining-based systems are more flexible and deployable.

Over the past several years, a growing number of research projects have applied data mining to intrusion detection with different algorithms. We propose an approach to use random forests and k-Nearest Neighbor in intrusion detection. For instance, those had been applied to prediction, probability estimation, and pattern analysis in multimedia information retrieval and bioinformatics.

Unfortunately, to the best of our knowledge, Random Forests algorithm has not been completely applied to detect novel attacks (unknown attacks) in automatic intrusion detection. Fortunately, we can take advantages from k-NN that can classify in more precisely and an important pattern recognizing method based on representative points.[2]

A. *Random Forest*

The Random Forests [4] is an ensemble of unpruned classification or regression trees. Random forest generates many classification trees. Each tree is constructed by a different bootstrap sample from the original data using a tree classification algorithm. After the forest is formed, a new object that needs to be classified is put down each of the tree in the forest for classification. Each tree gives a vote that indicates the tree's decision about the class of the object. The forest chooses the class with the most votes for the object.

The main features of the random forests algorithm are listed as follows:

- It runs efficiently on large data sets with many features.
- It can give the estimates of what features are important.

• It has no nominal data problem and does not over-fit.
• It can handle unbalanced data sets.

B. *k-Nearest Neighbor*

k-NN classification is an easy to understand and easy to implement classification technique[22]. Despite its simplicity, it can perform well in many situations. k-NN is particularly well suited for multi-modal classes as well as applications in which an object can have many class labels. For example, for the assignment of functions to genes based on expression profiles, some researchers found that k-NN outperformed SVM, which is a much more sophisticated classification scheme[2].

The 1-Nearest Neighbor(1NN) classifier is an important pattern recognizing method based on representative points [23]. In the 1NN algorithm, whole train samples are taken as representative points and the distances from the test samples to each representative point are computed. The test samples have the same class label as the representative point nearest to them. The k-NN is an extension of 1NN, which determines the test samples through finding the k nearest neighbors.

C. *Intrusion Detection System*

In this section, we describe the methods employed in the proposed system, and illustrate how to apply these methods to detect novel attacks with true positive rate, low false positive rate for network intrusion detection.

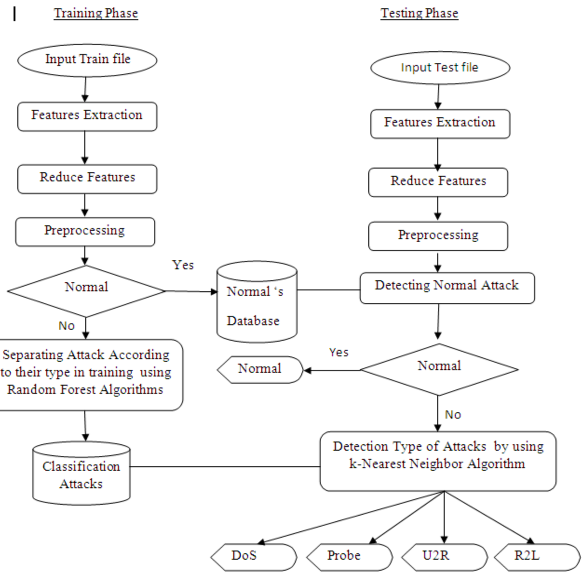

Fig 1. The proposed System

This system is process of identifying the abnormal and normal instances that are two phases. The first is the training phase that reduces the irrelevant features. Next phase is detection phase. This system is shown in Figure 1.

Since the operations of normal instances are specified and they show expected behavior, we could use the knowledge based (misuse) IDS detection, while unexpected activity (presumably an intrusion would be unusual) is continually designed and progressed and could not be seen as a knowledge based attack, therefore the anomaly IDS detection is performed over novel attacks.

We also report our experimental results over the KDD'99 datasets. The results show that the proposed approach provides better performance compared to the best results from the KDD'99 contest.

V. EXPERIMENTS AND RESULTS

In this section, we summarize our experimental results to detect unknown attacks for intrusion detection with over the KDD'99 datasets. Experimental results are presented in terms of the classes that achieved good level of discrimination from others in the training set.

Firstly, our proposed system will reduced some features in dataset by using Random Forest algorithm at each connection. So, system will try to detect various anomaly attacks using corrected KDD dataset. The proposed system will reduced in training time and will increase the accuracy of the system's classification. The experimental results will come out by using WEKA tool [15].

In the experiments process, the system use 10 trees and the reduced features (default 6 in WEKA) to classify. The accuracy of the system will be increased other systems as shown in Figure 2 and the detection rate using proposed method on each attack type are shown in figure 3.

Since the test datasets "Test+"and "Test-21" have with different statistical distributions than either "Train+" or "Train_20Percent", the accuracy decrease rather than Cross Validation results with those train files. But as to detect the unknown attack, the results in test file that contains more unknown attack types (novel attacks) than the other datasets get more detection rate of Random Forest can compare with other methods as shown in figure 2.

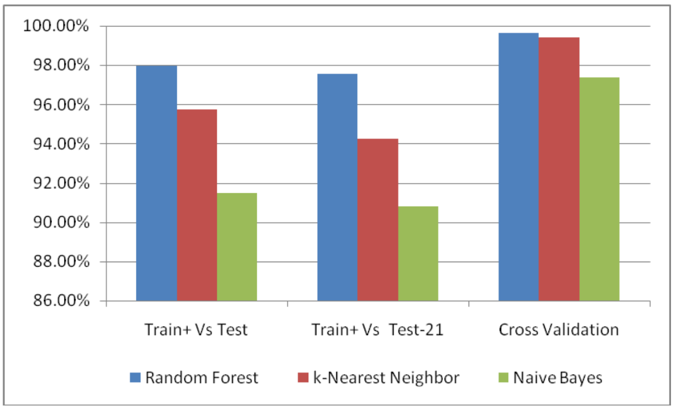

Fig. 2 The Comparison accuracy results between Machine Learning Algorithm Random Forest, k-NN and Naive Bayes.

VI. CONCLUSIONS AND FURTHER EXTENSION

Recent researches employed decision trees, artificial neural networks and a probabilistic classifier and reported, in terms of detection and false alarm rates, but it was still high false positives and irrelevant alerts in detection of novel attacks.

This paper has presented a survey of the various data mining techniques that have been proposed towards the enhancement of anomaly intrusion detection systems. And, we applied the classification methods for classifying the attacks (intrusions) on DARPA dataset. The results showing the performance of the Random Forest is better than other classifiers. But the time taken is more for Random Forest than other classifiers. On the other hand, k-Nearest Neighbor is also the good modeling algorithm in our experiments.The reason that the Random Forest cannot consider on pattern recognition, and also k-NN is a good pattern recognition method which used in many researches [3][21][22]. Thus, we can extend this experiment by combining those two algorithms; the system may expect to get the more accurate and detection rate to detected intrusion. Random Forest will process in the filtering stage and the k-NN will use as a classifier.

According to the experimental results and conclusion, we proposed a new model for more accurate and detection rate as shown in figure 3 using Knowledge Flow process in WEKA tools.

In this proposed model, as mention in conclusion, the Random Forest can process in feature ranking and selection in most research, we will used it in the filtering process of preprocessing state and it will construct the trees and also select the random features. After preprocessing state, we will use the k-NN algorithm, pattern recognition method for classification state to detect the incoming attacks.

Finally, we will drawn the results with text that express the Ture Positive, False Positive Rate, Precision, Recall and also confusion matrix we can extract.

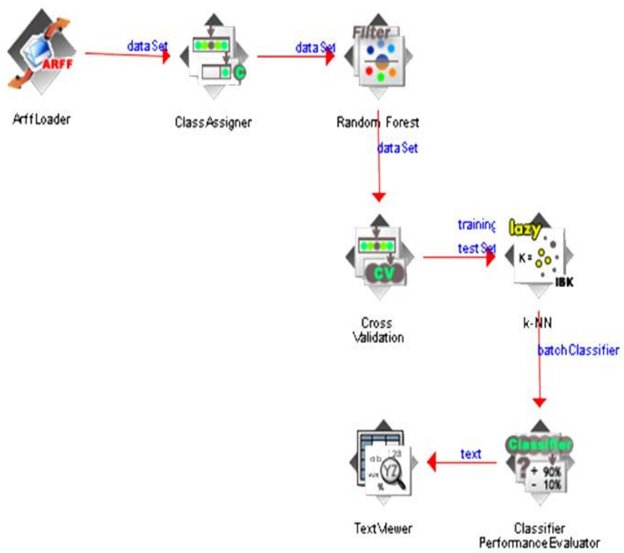

Fig. 3 The proposed Model

REFERENCES

[1] W. Lee and S. J. Stolfo, “Data Mining Approaches for Intrusion Detection”, the 7th USENIX Security Symposium, San Antonio, TX, January 1998.

[2] K.T.Khaing and T.T.Naing, “Enhanced Feature Ranking and Selection using Recirusive Featue Elemination and k-Nearest Neighbor Algorithms in SVM for IDS”, Internaiton Journal of Network and Mobile Technology(IJNMT), No.1, Vol 1. 2010.

[3] M. Bahrololum, E. Salahi and M. Khaleghi, "Anomaly Intrusion Detection Design using Hybrid of Unsupervised and Supervised Neural Network", International Journal of Computer Network & Communications(IJCNC), Vol.1, No.2, July 2009.

[4] L. Breiman, “Random Forests”, Machine Learning 45(1):5–32, 2001.

[5] V. Marinova-Boncheva, "A Short Survey of Intrusion Detection System" , 2007.

[6] Tamas Abraham, “IDDM: Intrusion Detection Using Data Mining Techniques”, DSTO Electronics and Surveillance Research Laboratory, Salisbury, Australia, May 2001.

[7] M. Mahoney and P. Chan, “An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection”, Proceeding of Recent Advances in Intrusion Detection (RAID)-2003, Pittsburgh, USA, September 2003.

[8] KDD’99 datasets, The UCI KDD Archive, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html , Irvine, CA, USA, 1999.

[9] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, December 2009.

[10] Lan Guo, Yan Ma, Bojan Cukic, and Harshinder Singh, “Robust Prediction of Fault-Proneness by Random Forests”, Proceedings of the 15th International Symposium on Software Reliability Engineering (ISSRE'04), pp. 417-428, Brittany, France, November 2004.

[11] Ting-Fan Wu, Chih-Jen Lin, and Ruby C. Weng, “Probability Estimates for Multi-class Classification by Pairwise Coupling”, The Journal of Machine Learning Research, Volume 5, December 2004.

[12] Yimin Wu, High-dimensional Pattern Analysis in Multimedia Information Retrieval and Bioinformatics, Doctoral Thesis, State University of New York, January 2004.

[13] Bogdan E. Popescu, and Jerome H. Friedman, Ensemble Learning for Prediction, Doctoral Thesis, Stanford University, January 2004.

[14] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Salvatore Stolfo. “A Geometric Framework for Unsupervised Anomaly Detection: Detecting Intrusions in Unlabeled Data.” Applications of Data Mining in Computer Security, 2002.

[15] WEKA software, Machine Learning, http://www.cs.waikato.ac.nz/ml/weka/, The University of Waikato, Hamilton, New Zealand.

[16] Leo Breiman and Adele Cutler, Random forests, http://statwww.berkeley.edu/users/breiman/RandomForests/cc_home.htm, University of California, Berkeley, CA, USA.

[17] David J. Hand, Heikki Mannila, and Padhraic Smyth, Principles of Data Mining, The MIT Press, August, 2001.

[18] MIT Lincoln Laboratory, DARPA Intrusion Detection Evaluation, http://www.ll.mit.edu/IST/ideval/,MA, USA.

[19] J.Zhange and M. Zulkerline, “Network Intrusion Detection using Random Forests”,2011.

[20] T. Lappas and K. Pelechrinis Data Mining Techniques for (Network) Intrusion Detection Systems”.

[21] J. Zhang and M. Zulkernine, ”Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection”, Symposium on Network Security and Information Assurance Proc. of the IEEE International Conference on Communications (ICC), 6 pages, Istanbul, Turkey, June 2006.

[22] S. Thirumuruganathan , “A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm”, World Press, May 17, 2010.

[23] X Wu, V Kumar, J Ross Quinlan, J Ghosh, “Top 10 Data mining Algorithm”, Knowledge and Information Systems, Volume 14, Issue 1, pp 1-37 ,2008 – Springer

[24] S. Mukkamala, A.H. Hung and A. Abraham, “Intrusion Detection Using an Ensemble of Intelligent Paradigms.” Journal of Network and Computer Applications, Vol. 28(2005), 167-182.

[25] S. Chebrolu, , A. Abraham, and J.P. Thomas, “Feature Deduction and Ensemble Design of Intrusion Detection Systems.” International Journal of Computers and Security, Vol 24, Issue 4,(June 2005), 295-307

[26] A.H. Sung and S. Mukkamala, “The Feature Selection and Intrusion Detection Problems.” Proceedings of Advances in Computer Science - ASIAN 2004: Higher- Level Decision Making. 9th Asian Computing Science Conference. Vol. 321(2004) , 468-482.

[27] S. Mukkamala, A.H. Sung and A. Abraham, “Modeling Intrusion Detection Systems Using Linear Genetic Programming Approach.” LNCS 3029, Springer Hiedelberg, 2004, pp. 633-642.

[28] A. Abraham and R. Jain, “Soft Computing Models for Network Intrusion Detection Systems.” Soft Computing in Knowledge Discovery: Methods and Applications, Springer Chap 16, 2004, 20pp.

[29] A. Abraham, C. Grosan, and C.M. Vide, "Evolutionar Design of Intrusion Detection Programs." International Journal of Network Security, Vol. 4, No. 3, 2007, pp. 328-339

[30] Shannon C.E., Weaver W., The Mathematical Theory of Communication. Urbana, IL, University of Illinois Press, 1946.

[31] Setiono R., Liu H., .Improving backpropagation learning with feature selection". Applied Intelligence: The International Journal of Arti_cal Intelligence, Neural Networks, and Complex Problem-Solving Technologies, vol. 6, pp. 129-139, 1996.

[32] Press, W.H., Flnnery B.P., Teukolsky S.A. and Vetterling W.T. .Numerical recipies in C.. Cambridge University Press, Cambridge.

[33] Chi J., .Entropy based feature evaluation and selection technique., Proc. of 4th Australian Conf. on Neural Networks (ACNN'93), pp. 181-196, 1993.

Phyu Thi Htun received the B.E. degree in Information and Technology Engineering from Government Technological College, Thanlyin , Myanmar, in 2006. And she received M.E degree in Information and Technology Engineering from Westen Yangon Technology University,Yangon,Myanmar, in 2010 respectively. She is currently doing research for Ph.D(IT) in University of Technology ,Yadanapon Cyber City , Myanmar since November 2010. Her research interests are in network security, especially intrusion detection systems, and survivability of cloud computing.

Kyaw Thet Khaing received the Master of Computer Technology from University of Computer Studies,Mandalay , Myanmar, in 2004. And he received Ph.D degree in Information Technology from University of Computer Studies,Yangon,Myanmar, in 2010 respectively. His research interests are in network security, especially intrusion detection systems, and survivability of cloud computing.