Three-level Hybrid Intrusion Detection System

Hui Lu
Department of Computer Science
East China Normal University
Shanghai, China
luhui5599@126.com

Jinhua Xu
Department of Computer Science
East China Normal University
Shanghai, China
jhxu1008@yahoo.com

***Abstract*—With increasing connectivity between computers, the need to keep networks secure becomes more and more vital. Intrusion detection systems have become an essential component of network security to supplement existing defenses. This paper proposes a novel intrusion detection system, which combines the supervised classifiers and unsupervised clustering to detect intrusions. Decision Tree, Naïve Bayes and Bayesian clustering are used at different levels. We also have made improvements to the Naïve Bayes algorithm by choosing different attributes for different classes. The experiments demonstrate the effectiveness of the proposed approach, especially for U2R and R2L type attacks. The detection rate is significantly improved.**

Keywords- Intrusion detection;internet security;Decision Tree;Naive Bayes;Bayesian Clustering

I. INTRODUCTION

With the popularity of internet, internet affects the polity, economy, culture, military and life. There are more and more issues settled on the internet, such as Email, banking, video conference and so on. Therefore, the internet security becomes one of the key problems in the world. According to the statistics, the global number of virus has been more than 40,000. The average network intrusion incident occurs every twenty seconds. Each year the global loss caused by the issue of security can be calculated in the magnitude of one trillion U.S. dollars. Information security directly influences the interests of nation, corporation and individual. Traditionally, firewall is widely used security measures. Nowadays, the tools and tactics turn more complex and diverse. A simple firewall policy has been unable to meet people's needs. Therefore, the protection of computer systems, network systems and the security of information infrastructure have to be addressed immediately. Intrusion detection technology is an important part of internet security.

Intrusion detection is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a system or network. An intrusion detection system is a software tool used to detect unauthorized access to a computer system or network [1].

So far, many techniques for intrusion detection have been reported. The techniques for intrusion detection are traditionally classified into two categories: misuse detection and anomaly detection [2]. Misuse detection techniques try to model attacks on a system as specific patterns, and then systematically scan the system for occurrences of these patterns [3]. Anomaly detection is an approach to detect intrusions by first learning the characteristics of normal activity. Then systems are designed to detect anything that deviates from normal activity [4]. However, misuse detection can not detect some known attacks occasionally. Anomaly detection suffers a higher false rate.

Thus, hybrid detection systems have been proposed to deal with this problem, which employs both misuse detection and anomaly detection. For example, EMERALD [5] is a hierarchical intrusion detection system. Depren etal [6] proposed an intelligent intrusion detection system making use of both anomaly and misuse detection. Zhang and Zulkernine[7] raised a hybrid intrusion detection system using random forests. The other type of hybrid detection systems used different machine learning algorithm. Pfahringer [8] utilized bagged boosting algorithm. Levin [9] suggested a KDD-99 classifier to deal with the intrusion detection by using kernel miner. Peddabachigari etal [10] applied support vector machines and decision tree. While Cheng Xiang et al [11] put forward a multiple-level tree classifier. This approach has a high false alarm rate and a low detection rate for U2R and R2L attacks. Then on this basis, they raised a new multiple-level hybrid classifier using Bayesian clustering and decision trees [12]. Although the latter system is more effective, the detection rates of U2R and R2L are still low. In this paper, we proposed a novel three-level hybrid intrusion detection system to solve this problem. The detection rates of U2R and R2L type attacks are improved, as shown in the experiments.

The rest of the paper is organized as follows. Section 2 introduces Decision Tree, Naïve Bayes and Bayesian clustering algorithms. Section 3 gives details of the novel intrusion detection model. Section 4 presents the experimental results; Conclusions and future work are discussed in section 5.

II. ALGORITHM

This section introduces the algorithms used in the model. We briefly discuss the C4.5 and Bayesian Clustering algorithms. The novel Naïve Bayes algorithm is described in detail.

A. C4.5 Algorithm

C4.5 algorithm is a later version of the ID3 algorithm [13]. ID3 is designed for the scope, where there are many attributes and the training set contains many objects, but where a reasonably good decision tree is required without much

978-1-4244-4994-1/09/$25.00 ©2009 IEEE

computation [14]. The process of tree building continues until all the leaf nodes are pure or there are no other branch variables. The essence of ID3 is iterative. The computation of impurity is estimated in [15]

$$\Delta i_B(s) = \frac{\Delta i(s)}{-\sum_{k=1}^{B} P_k \log_2 P_k} \quad (1)$$

When $\Delta i_B(s)$ is maximum, the branch is optimal. The normal ID3 is unable to do pruning operation. While the C4.5 algorithm uses heuristic techniques to achieve pruning. In the stage 1, we use this algorithm to built decision tree. The detail about Decision Tree has been introduced in [13]-[15].

B. Bayesian Clustering Algorithm

Bayesian clustering is also called AutoClass[16]. It utilizes finite mixture model and Bayesian method for determining the optimal clusters. In order to find clusters in a dataset, the number of cluster should be initialized with an assumption achieved from experience. Then, the task of classification becomes to estimate these classification parameters from a given dataset. The EM algorithm has been used to estimate the parameters of the probability distributions to best fit the data. But there is no guarantee that the EM algorithm converges to the global optimum. The procedure is repeated for several different sets of initial values. AutoClass also considers different numbers of clusters and different amounts of covariance and different underlying probability distribution types for the numeric attributes [17]. The overall algorithm is extremely time-consuming. In fact, the actual implementation starts with a pre-specified time bound and continues to iterate as long as time allows. Give it longer and the results may be better

C. Naive Bayes Algorithm

Many classifiers can be viewed as computing a set of probability distribution functions of the example, one for each class, and assigning the example to the class whose probability is maximum [18]. In the data processing, Naïve Bayes assumes that the attributes are completely independent. If some attributes are dependent. Naïve Bayes does not do well. We can avoid this problem by a careful selection of the attributes to be used. The normal-distribution assumption for numeric attributes is another restriction on Naïve Bayes. Many attributes are not normally distributed. In this case, the performance of Naïve Bayes is poor. In this paper, kernel density estimation is used to deal with this issue.

A record is defined by n attributes, namely $X = (x_1, x_2, ..., x_n)$. Suppose that there are m categories $C_1, C_2, ..., C_m$. We calculate $P(C_i \mid X)(i = 1, 2, ..., m)$ and select the maximum of $P(C_i \mid X)$. Then, the record x is classified into category C_i . $P(C_i \mid X)$ is posterior probability and defined by

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \quad (2)$$

Because $P(X)$ is a constant, we only calculate $P(X \mid C_i)P(C_i)$ and select the maximum value. In Naïve Bayes, we assume the attributes are independent, that is,

$$P(X \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i)$$

The probability of $P(x_1 \mid C_i), P(x_2 \mid C_i), ..., P(x_n \mid C_i)$ can be evaluated from the training set. Specific methods are introduced as follows.

1) For discrete attribute:

$$P(x_k \mid C_i) = \frac{N(x_k \mid C_i)}{N(C_i)} \quad (3)$$

Where $N(C_i)$ is the number of samples in class C_i , $N(x_k \mid C_i)$ is the number of samples in class C_i with attribute value x_k . If $N(x_k \mid C_i) = 0$ in the training set, then set $P(x_k \mid C_i) = 0.0001$.

2) For continuous attribute: In this paper, Gaussian kernel method is used to estimate the probability density function of a continuous variable [19].

$$P(x_k \mid C_i) = (n_{C_i}\sigma)^{-1} \sum_{x_k : C(x_k) = C_i} K\left(\frac{x_k - \mu_i}{\sigma}\right) \quad (4)$$

Where

$$K(x) = g(x, 0, 1) \quad (5)$$

And

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\pi)^2}{2\sigma^2}} \quad (6)$$

If the probability is zero, set it to 0.0001.

Generally all models choose the same attributes. In our algorithm, models for different classes choose different attributes. In the stage 2, the U2R model and the hybrid model of R2L and Normal are established. The former uses all attributes, while the latter selects 38 attributes. The attributes of dst_host_srv_count, protocol_type and service are not used in the latter model. Experiments have shown this improvement is positive.

III. THREE-LEVEL HYBRID CLASSIFIER MODEL

This section introduces the model used in the experiment. The architecture of the three-level hybrid classifier is shown in Fig 1. In stage 1, Decision Tree Algorithm is used to separate records into three categories----DOS, Probe and Other. The type of Other includes U2R, R2L and Normal categories. The main purpose of this stage is to extract U2R and R2L and Normal records. In stage 2, it categorized the type of Other into two categories-----U2R, Rest. Normal and R2L are classified as Rest. Judging from the other experiment results, the detection rates of U2R and R2L are still low. So, it is important to enhance the detection rate of the two types. A large number of experiments show that Naïve Bayes Algorithm can improve the detection rate of U2R and can extract as many Normal and R2L records as possible from the data. The stage 3 separates Rest into Normal and R2L categories. In this stage, Bayesian Clustering is used. It is mainly to improve the detection rate of the R2L.

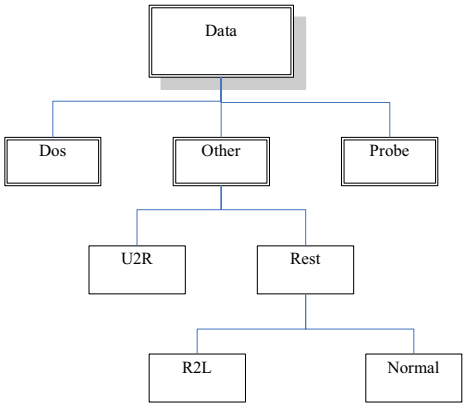

Fig 1. Architecture of the three-level hybrid classifier

Compared with the model in [12], stage 1 is same, but stage 2 and stage 3 are different. In [12], Stage 2 separates out the attack (including U2R and R2L) and Normal connections using Bayesian clustering; Stage 3, decision trees are used to separate out the U2R and R2L.

IV. EXPERIMENTS AND RESULTS

This section describes the training and test data used in the experiment, the experimental procedures and results.

A. *Training and Test data*

The dataset used in the experiment comes from KDD CUP 99 dataset, which is used for The Third International Knowledge Discovery and Data Mining Tools Competition [20]. KDDCUP99 dataset is an extension of DARPA98 dataset with a set of additionally constructed features [21]. Each connection record is defined by 41 features and one more feature assigns the record type. The datasets contain a total of 22 training attack types, with an additional 17 types in the test data only. Table 1 shows KDD CUP 99 dataset attack types and patterns [22].

In the experiment, the training data selected from the full data set of KDD CUP 99 contains 22 attack types. It is more than 20,000 records. The test data utilizes all of the test data in KDD CUP 99, totally 311029 records.

TABLE I. KDD CUP 99 DATASET PATTERN AND CLASSIFICATION

Attack type	Attack pattern
Probe	ipsweep, nmap, portsweep, satan, mscan, saint
Dos	back, land, neptune, pod, smurf, teardrop, apache2, mailbomb, processtable, udpstorm
U2R	buffer_overflow,loadmodule,perl, rootkit,ps,sqlattack,xterm
R2L	ftp_write ,guess_passwd,imap,phf, multihop,spy,warezclient,warezmaster,snmp, getattack,named,xlock, xsnoop,snmpguess,worm,httptunnel,sendmai l

B. *Software*

Weka written in java was developed at the University of Waikato in New Zealand. It provides a uniform interface to many different learning algorithms, along with methods for pre- and post-processing and for evaluating the result of learning schemes on any given dataset [17], [23]. In stage 1, the experiment in this paper used J48 algorithm which was included in weak 3.5.8. J48 algorithm carries out C4.5 decision tree learner.

AutoClass is an unsupervised Bayesian classification system. The version used in the experiment is Autoclass-c-3-3-4. [24].

C. *Experimental procedures*

The full data set of KDD CUP 99 is less than five million. We select about two hundred thousand records as training set and use all attributes. In stage 1, first of all, pre-process data fit for Weka. Then, J48 algorithm is used to establish decision tree model. We can set the parameters of J48 algorithm, such as confidenceFactor, numFolds, uselaplace, etc.

In stage 2, the U2R, R2L and Normal records are separated out with a novel Naïve Bayes algorithm, which is described in section 2. We extracted the Normal, R2L and U2R records from KDD CUP 99 dataset. The experiment used all of the records to built Naïve Bayes classifier. Besides, all the examples in this section are executed using Matlab.

The last stage, Bayesian clustering algorithm is used to split Rest into Normal and R2L categories. We adopt the method in Cheng Xiang's experiment [12]. That is to say, the clusters which contain at least one R2L attack are all labeled 'R2L'.The training data used is ten thousand normal records and all of the R2L attacks from KDD CUP 99 dataset. We selected four features for clustering. The four attributes selected are duration, service, src_bytes and dst_bytes. This stage we used AutoClass C [24]. In our experiment, the initial number of clusters is fixed at 200. At the end of experiment, the training set is divided into 185 clusters.

D. *Experimental Results And Discussion*

The detection rates for the five categories compared to other methods are shown in table 2. In this paper, the detection rate is the ratio of the number of correct detection to the total number of each type.

From table 2, the detection rate of DOS and Probe is similar to other methods. The detection rate of these two types is still good. It is obvious that the detection rate of U2R is very high.

TABLE II. COMPARISON OF DETECION RATES (%)

Type	Bagged boosted c5 trees	Kernel miner	Three-Level tree	Multiple-level hybrid	Three-level hybrid
Dos	97.10	97.47	97.35	98.66	98.54
Probe	83.30	84.52	93.23	93.40	93.50
U2r	13.2	11.84	61.43	71.43	97.14
R2l	8.40	7.32	23.69	46.97	48.91
Normal	99.50	99.42	42.73	96.80	94.68

It reaches to 97.14%. The improvement of detecting U2R over multiple-level hybrid classifier is significant. The detection rate of R2L is higher than other methods. Although the false alarm rate is 5.32%, it is still acceptable. If attack records are misclassified as normal records, it may cause great harm to the internet. While if normal records are labeled attack, it may only bring inconvenience to the network administrator. From this point of view, the three-level hybrid classifier makes great progress on improving the detection rate of attack types, especially U2R. Lots of experiments have showed that when the detection rate of R2L amounts to 60%, the false alarm rate will be as high as 40%. In other words, the three-level hybrid classifier proposed in this paper can get a high detection rate with an acceptable false alarm rate. Thus, it is proved to be effective.

V. CONCLUSION AND FUTURE WORK

The three-level hybrid intrusion detection system utilized Decision Tree, Naïve Bayes and Baysian Clustering. The training and test data used in this paper is KDD CUP 99 dataset. Experiment results compared with other representative methods, such as kernel miner and multiple-level hybrid classifier. The experiment results using the proposed model show that the detection rate of U2R and R2L is much higher than others, especially U2R. The detection rate of U2R is 97.14%, while the detection rate obtained by the Bagged boosted C5 trees is only 13.2%. The Multiple-level hybrid classifier can detect U2R attacks with 71.43%. For R2L, the detection rate is 48.91%. While the detection rate got by the three-level tree classifier is only 23.69% and the false alarm rate is 57.27%.

Although the three-level hybrid classifier is suitable, there is still much room to improve the detection rate of R2L. Our goal is to seek higher detection rate and lower false alarm rate. The model has great impact on the detection rate. Using different pattern recognition algorithm can get different detection rate for each type data. For example, the detection rate of DOS is extraordinarily high with the Decision Tree Algorithm. In the experiment, we find that Naïve Bayes can separate out the U2R and R2L records well. If we only use these two types of records, the detection rate of U2R and R2L is 97.14% and 93.82%. This result is premised on the fact that U2R and R2L can extract from the dataset well. However, it is difficult to achieve this premise. If the goal is accomplished, the detection rate of attack categories will be greatly improved, especially U2R and R2L. In further studies, we will strive for new models and more suitable pattern recognition algorithm, and make a breakthrough in the study of the intrusion detection.

REFERENCES

[1] Animesh Patcha , Jung-Min Park. “An overview of anomaly detection techniques: existing solutions and latest technological trends”, Computer Networks 51, pp.3448-347, 2007

[2] D. Anderson, T. Frivold, A. Valdes, “Next-generation intrusion detection expert system (NIDES): a summary”, SRI-CSL-95-07, May 1995.

[3] A. Ghosh, A. Schwartzbard, “A study in using neural networks for anomaly and misuse detection”, in: Proceedings of the Eighth USENIX Security Symposium, Washington, pp. 141−152, August 1999..

[4] Kemmerer, R. A., & Vigna, G., “Intrusion detection: A brief history and overview”, IEEE Security and Privacy Magazine, 2002.

[5] P.A. Porras, P.G. Neumann, “EMERALD: event monitoring enabling responses to anomalous live disturbances”, in: Proceedings of the 20th NIST-NCSC National Information Systems Security Conference, Baltimore, MD, USA, pp. 353−365, 1997..

[6] Ozgur Depren, Murat Topallar, Emin Anarim, M. Kemal Ciliz, “An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks”, Expert Systems with Applications, 29 ,pp.713–722, 2005.

[7] Zhang, J., Zulkernine, M., “A hybrid network intrusion detection technique using random forests”, In: Proc. 1st Internet. Conf. on Availability, Reliability and Security, ARES, Vienna, Austria, pp. 262−269 , 2006

[8] Pfahringer, B., “Winning the KDD99 classification cup: Bagged boosting”, SIGKDD Explor, 1 (2), pp.67−75, 2000..

[9] Levin, I., “KDD-99 classifier learning contest LLSoft ’ s results overview”, SIGKDD Explor. ACM SIGKDD, 2000..

[10] Peddabachigari, S., Abraham, A., Grosan, C.,Thomas, J., “Modelling intrusion detection system using hybrid systems”, J. Network Comput. Appl. 30, pp.114−132, 2007.

[11] Xiang, C., Chong, M.Y., Zhu, H.L., “Design of multiple-level tree classifiers for intrusion detection system”, In: Proc. 2004 IEEE Conf. on Cybernetics and Intelligent Systems, Singapore, pp. 872−877, December 2007.

[12] Cheng Xiang,Png Chin Yong, Lim Swee Meng, “Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees”, Pattern Recognition Letters, 29, pp.918-924, 2008..

[13] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993..

[14] J.R. Quinlan, “Introduction of decision trees”, Machine Learning, 1 , pp81– 106, 1986..

[15] Richard O.Duda Peter E. Hart David G. Stork,Pattern Classification,Second Edition, John Wiley&Sons,inc., 2001..

[16] Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., Freeman, D., “AutoClass: A Bayesian classification system” In: Proc. Fifth Internat. Conf. on Machine Learning,1988..

[17] Ian H.Witten, Eibe Frank, Data Ming: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers. ISBN: 0-12-088407-0..

[18] Duda, R. O., & Hart, P. E.. Pattern classification and scene analysis. New York, NY: Wiley,1973..

[19] George H. John, Pat Langley, “Estimating Continuous Distributions in Bayesian Classifiers”, in proceedings of the eleventh conference on uncertainty in artificial intelligence, Morgan Kaufmann Publishers, San Mateo,1995.

[20] KDD Cup, 1999. Data, Information and Computer Science, University of California, Irvine. [Online]. Available: http://KDD.ics.uci.edu/databases/KDDcup99/KDDcup99.html.

[21] Lee, W., Stolfo, S.J., 2000, “A framework for constructing features and models for intrusion detection systems”, ACM Trans. Inform. Syst. Security 3 (4), 227−261..

[22] Wu, S.-Y., & Yen, E., “Data mining-based intrusion detectors”, Expert Systems with Applications,doi:10.1016/j.eswa.2008.06.138.

[23] Weka 3: Data Mining Software in Java, University of Waikato, New Zealand. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka.

[24] Obtaining AutoClass C. [Online]. Available: http://ic.arc.nasa.gov/ic/projects/bayes-group/AutoClass/AutoClass-c-program.html...