Research on Intrusion Detection Model Using Ensemble learning Methods

Ying Wang, Yongjun Shen and Guidong Zhang
School of Information Science and Engineering
Lanzhou University
Lanzhou, China
shenyj@lzu.edu.cn

***Abstract*— In the past few years, for security issues keep increasing and changing explosively, Intrusion Detection System (IDS) have been asked for discovering these threats more accurately and more rapidly. Many intrusion detection models have been proposed to meet the requirements. However, for this unbalanced data sample KDDcup99, some classifiers may have a good effect on big sample classes like J48 and RandomForest, others are good at classifying on small sample, but few of them can achieve balance. In this paper, an intrusion detection solution based on ensemble learning is put forward. In our model, we employ Bayesian network and RandomTree as base classifiers along with meta learning algorithms RandomCommitte and vote. To evaluate the model's performance, the KDDcup99 dataset is introduced.**

Keywords—Intrusion detection; ensemble learning; Bayesian networks; RandomTree; meta learning ;KDDcup99

I. INTRODUCTION

Although the cyber-attacks emerge in endlessly and diversely, IDSs still play an important role in defending network security. With the development of intrusion technologies, IDSs need make adjustments to adapt for changes in security requirements. Traditional IDSs can be divided into anomaly detection and misuse detection. These IDSs have advantages in discovering some behaviors which are harmful for information systems and network, but also have their shortcomings in detection process. Anomaly detection based on modeling the normal activity of the computer system is general enough to detect new attacks with low false alarm rates [1]. However, anomaly detection can result in high false positives. Misuse detection is an attack signature-based approach that utilizes a detailed description of the sequence of actions performed by the attacker. More general signatures would reduce these misdetections but also give high false alarm rates [1].

Both misuse detection and anomaly detection can be regarded as classification and modeling process, for which, there is no difference but modeling object. With the widespread use of data mining technology, many investigators have combined intrusion detection modeling with these technologies, such as support vector machine (SVM), neural network, logistic regression (LR) and so on. Some of their models have achieved good results on experiments. However, there still is some space for us to study, because of the unbalanced dataset (which we will mention in the following sections) and algorithms' inherent flaws. In our investigation, we adopt ensemble learning to build intrusion detection model, Bayesian network and RandomTree are employed as base classifiers.

The paper is organized as follows. In section II, we introduce some related work about intrusion detection models using data mining methods; in section III, we give a detailed description about our method for building intrusion detection model; in section IV, we test our model using the kddcup99 dataset and compare to other methods; in section V, we draw a conclusion for this paper.

II. RELATED WORK

In general, several approaches can be used for improving intrusion detection performance, and one of these is classification along with feature selection. XIAO et al give a method called RS-MSVM (Rough Set and Multi-class Support Vector Machine) which is proposed for network intrusion detection. This method is based on rough set (RS) followed by MSVM for attribute reduction and classification respectively [2]. Similarly, another paper also uses RS for feature selection. What different is that dataset classified by SVM using bagging [16]. Hamid Ghaffari Gotorlar et al present a solution, in their study, harmony search (HS) with a novel normalization method is embedded in the SVM model to make HS-SVM so as to shorten the testing time and improve the performance of SVM [3]. Also, we can utilize Principle Component Analysis for feature selection. a combined algorithm based on Principal Component Analysis (PCA) and Core Vector Machine (CVM), which is an extremely fast classifier, is proposed for intrusion detection[4]. Another paper also adopt PCA for feature selection, instead of the CVM algorithm, they classify by Artificial Neural Networks using a meta learning algorithm called randomcommittee [5].

Undoubtedly, some different ideas have been achieved. Two statistical methods viz. Linear Discriminant Analysis (LDA) and Logistic Regression (LR) are applied to develop new intrusion detection models [6]. Thulasy Ramiah Pillai et al introduce Generalized Autoregressive Moving Average (GARMA) and Autoregressive Moving Average (ARMA) into their model [7]. A concept called complex machine learning (CML) is proposed, this model use hybrid-classifiers which is effective for large dataset and unknown attacks [8]. In [1], they present dLEARNIN, which utilizes an ensemble of classifiers approach that combines information from different sources of

978-1-4673-9904-3/16/$31.00 ©2016 IEEE

information. Also, they utilize a cost minimization strategy dCMS, to gear the final classification decision toward minimizing the cost of the errors, the true objective function, and not the error rate itself. In [9], a new solution called fast inductive learning applied in intrusion detection model. This fast inductive learning method for intrusion detection (FILMID) improved RIPPER which is traditional rule-based inductive learning algorithm developed by Cohen of AT&T lab, and gain good effect.

Models based on Bayesian network for anomaly detection and misuse detection as we mentioned in section I are advised by Wojciech Tylman [10, 11]. In [12], they build a Bayesian classifier by Bayesian Model Averaging (BMA) over the k-best BN classifiers, called Bayesian Network Model Averaging (BNMA) classifier. There also are some investigations using ensemble learning for intrusion detection. In [13], ensemble learning using adaboost is adopted for detecting intrusion behaviors. In addition, correlation-based algorithm is used for reducing some redundant features. C5.0 classification using adaboost for intrusion detection is presented [18]. In [17], incremental GHSOM is applied in Intrusion detection.

In [14, 15], we find Bayesian network and RandomTree performs better than other classifiers in some aspects. However, when we train and test our pretreated data using one of these alone, the results can't meet what we expected. So ensemble learning methods are considered for raising their performance.

III. METHODS

In our study, ensemble learning is imported for intrusion detection model. As we all know, ensemble learning can be achieved in several ways as follows

1). Using the same classifier, but different in samples.

2). Using the same data sample, but different in classifiers.

3). Making use of sample's different attributions for building base classifiers

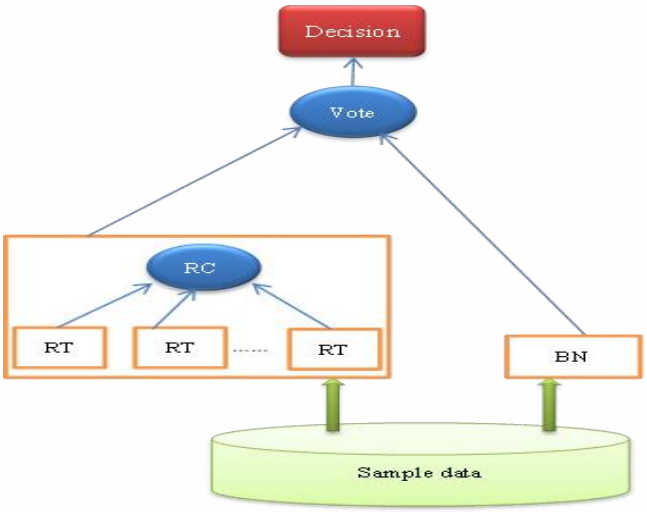

Fig. 1. Intrusion detection model architecture

Here we adopt the first two methods for ensemble learning. Bayesian network combines RandomTree are employed for intrusion detection. First of all, we use a meta learning algorithm called RandomCommittee (RC) raising RandomTree (RT) as one of base classifiers, then this classifier and Bayesian network (BN) are voted for classifying the dataset which we have preprocessed. As we can see in figure 1

A. Dataset

For evaluating the performance of the model we put forward more objectively and fairly, KDDcup99 dataset developed by MIT Lincoln Lab[4] is introduced. In view of the enormous of this dataset which include about 5 million records, we resample it's 10% dataset which contain 494021 records. There are 22 attacks that can be divided into 4 types (Probe, DOS, U2R, R2L) in this KDDcup99' 10% data. We can see from the table 1, U2R and R2L only are very small parts of the whole dataset. As for the dataset exits too many repeat records which make no sense for our experiment, so we remove these duplicates and get our final training and test dataset as table 1 shows.

TABLE I. EXPERIMENTAL DATA SETS

	KDDcup_10% dataset	Dataset_remove duplicates
Normal	97278	87832
Probe	4107	2122
DOS	391458	54156
U2R	52	52
R2L	1126	999

B. Bayesian network

For some of their excellent characteristics, Bayesian network classifications have been widely applied in many areas, such as pattern Recognition, disease diagnosis, information retrieval and so on. Bayesian network are composed of directed acyclic graph (DGA) and set of conditional probability distribution (CPD) of each child node of DGA.

In DGA, nodes represent random variables which can be continuous or discrete. Edges between node and node represent relations (casual or probability relationship) between variable and variable. When one variable is discrete, CPD can be instead of conditional probability table (CPT). It is worth noting that conditional probability have different senses in Bayesian network compared to Navies Bayes. In Navies Bayes, properties are conditional independent or independent. However, there are many correlations between properties and properties in Bayesian network. Providing the set of random variables is $X=\{x_1,x_2,\ldots,x_n\}$, in general, the joint probability conditional distribution can be presented as

$$P(x_1, x_2, \ldots, x_n) = P(x_1)P(x_2|x_1)P(x_3|(x_1,x_2)) \ldots P(x_n|(x_1,x_2,\ldots,x_{n-1})) \quad (1)$$

Because each node of DGA is conditional independent form its non-parents nodes, so the joint probability conditional distribution can be presented as

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|(Parents(x_i)) \quad (2)$$

Building and training of Bayesian network classifier can be divided into two steps as follows:

1) We need determine the topology relationship between random variables and form DAG, which usually require experts in the field to complete. However, for establishing a good topology usually requires constant iteration and improvement.

2) When the first step has been finished, we train the Bayesian network. This step is to complete the construction of conditional probability table, if the value of each random variable is directly observable, then this step training is intuitive, a method similar to Naive Bayes classifier. But usually Bayesian networks exist hidden variable nodes, the training method is more complicated, such as gradient descent methods.

C. *RandomTree*

The randomtree is one of tree classify algorithms which different from general decision tree classifications. Not all but a part of attributions can be selected randomly by RandomTree classifier for building child trees. Generally, RandomTree converge faster than other decision trees like J48, and performance of classification behaves not weaker than others'. So in this model we choose RandomTree as one of base classifiers. We can describe the attribution selection and split points generate process of RandomTree like one blog [19] said.

1) Set a number K stands the mount of attributions to be selected.

2) Resample attributions without replacement in the whole attributions.

3) Compute information gain of the attrbution

4) Repeat this K times and select the splitting node whose maxinum information gain.

5) Build child trees of the split node.

D. *RandomCommittee & vote*

RandomCommittee is one of meta learning algorithms, who can integrate RandomTree or RandomForest and any other randomizable classifiers as its base classifiers. The method of RandomCommittee raising base classifiers is similar to the first method we mentioned about ensemble learning. What different is that each base classifier use the same dataset, but different seed in per base classifier [5].

Vote is another meta learning algorithm who can also optimize base classifiers' classification performance. Vote can combine different classifiers, usually classifiers combine using vote gain better performance than these alone. There are several combination rules for vote, such as average of Probabilities, majority of vote, product of probabilities, maximum of probabilities and so on. Here we adopt average of probabilities as combination rule for vote.

IV. RESULTS& EVALUATION

In this part, we will utilize KDDcup99 dataset we mentioned in section III and 10-fold cross-validation for testing our intrusion detection model and getting results. Then we will evaluate performance of the model. In classification issues, providing that the correctly classified mount of one class is TP, correctly classified mount of other classes is TN, the mount of false classified into this class is FP, the mount false classified into other classes is FN, then we can define accuracy rate of the whole dataset as follows:

$$\text{Accuracy rate} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{TN} + \text{FN}) \quad (3)$$

For evaluating algorithms in the view of classification, accuracy rate can't be ignored. Figure 4 presents the accuracy rate of each attack for three classifiers. We can get from this figure that our intrusion detection model has a higher accuracy for per attack classes when compare with Bayesian network and RandomTree.

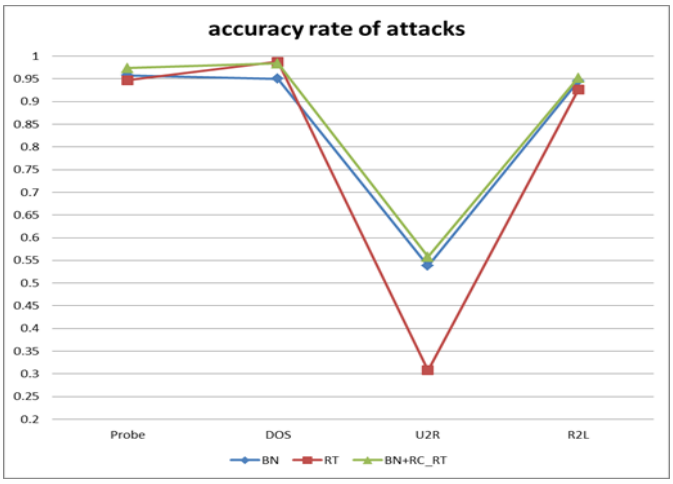

Fig. 2. Accuracy rate of attacks for each model

Then we evaluate our model using receiver operating characteristic (ROC) curves which response the sensitivity and specificity of classification models. We paint ROC curves of these attacks (Probe, DOS, U2R, R2L) respectively in figure 3, figure 4, figure 5, figure 6. For ROC curves, which curve nearest to the upper left corner of the figure indicates have the best performance. According to these figures, we can infer that our intrusion detection model has higher performance when compared to Bayesian network and RandomTree on sensitivity and specificity.

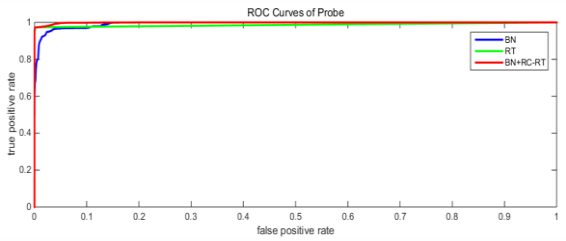

Fig. 3. ROC Curves of Probe attacks

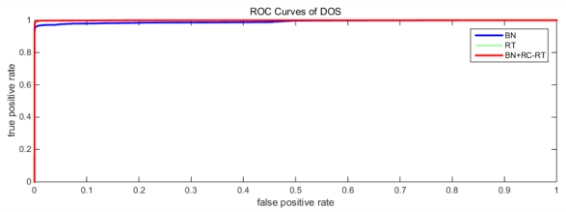

Fig. 4. ROC Curves of DOS attacks

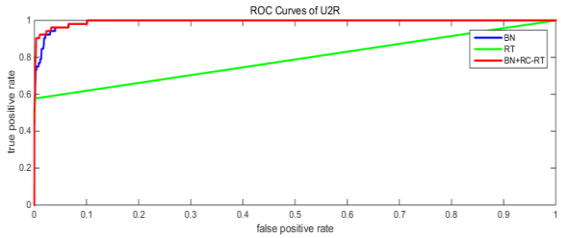

Fig. 5. ROC Curves of U2R attacks

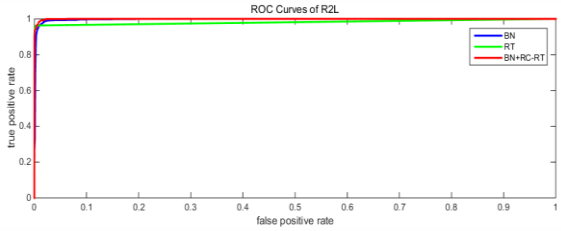

Fig. 6. ROC Curves of U2R attacks

In order to gain a more intuitive and quantized consequence on sensitivity and specificity, we can calculate the area under the ROC curves (AUC). When compared to Bayesian network and RandomTree, our model has a bigger area for each attack, which also means a higher performance.

TABLE II. AUC OF ATTACKS ON PER MODEL

Classifier	Probe	DOS	U2R	R2L
BN	0.993	0.991	0.992	0.997
RT	0.987	0.999	0.788	0.981
BN+RC_RT	0.999	1.000	0.995	0.999

V. CONCLUSION

In this paper, we put forward one intrusion detection model based on Bayesian network and RandomTree using ensemble learning methods. Then for evaluating our model, we run this model on KDDcup99 dataset which we have pretreated. We compare our model with base classifiers alone from accuracy rate, sensitivity and specificity. For sensitivity and specificity, we adopt ROC curves and AUC to evaluate the model we proposed. The consequence shows that our model has a better effect on sensitivity and specificity. In terms of accuracy, base classifiers have their pros and cons. Bayesian network presents higher accuracy for small sample classes, but have lower accuracy rate than other classifiers. On the contrary, RandomTree present good performance for big sample but not good at small sample. Our model using ensemble learning combines advantages of Bayesian network and RandomTree and presents a good effect for the whole dataset whatever big sample or small sample.

Although the performance of our model has been improved, there are some other aspects we can concentrate on. For example, as for U2R attacks, accuracy rate still is too low which hard to be applied in the actual intrusion detection activities. In order to continue improve accuracy rate and detection rate of small sample attack classes, cost-sensitive learning methods can be considered. In addition to cost-sensitive learning methods, other solutions aimed at unbalanced dataset like SMOTE also can be introduced. In the future work, we will carry on completing our model from these methods.

REFERENCES

[1] Parikh, D. and T. Chen, Data Fusion and Cost Minimization for Intrusion Detection. IEEE Transactions on Information Forensics and Security, 2008. 3(3): p. 381-389.

[2] Xiao Yun , H.C., Zheng Qinghua ,Zhang Junjie, NETWORK INTRUSION DETECTION METHOD BASED ON RS-MSVM. JOURNAL OF ELECTRONICS, 2006. 23: p. 901-905

[3] Hamid Ghaffari Gotorla, J.B., Mohammad Pourmahmood Aghababa, Masoumeh Samadi Osalu, Improving Intrusion Detection Using a Novel Normalization Method along with the Use of Harmony Search Algorithm for Feature Selection, in 7th International Conference on

[4] P.Amudha, S.K., S.Sivakumari, Intrusion Detection Based on Core Vector Machine and Ensemble Classification Methods, in International Conference on Soft-Computing and Network Security. 2015, IEEE.

[5] Milde M. S. Lira, R.R.B.d.A., Aida A. Ferreira, Manoel A. Carvalho Jr, Otoni Nóbrega and G.S.M.S. Neto, Combining Multiple Artificial Neural Networks Using Random Committee to Decide upon Electrical Disturbance Classification, in Proceedings of International Joint Conference on Neural Networks. 2007: Orlando, Florida, USA.

[6] Basant Subba, S.B., Sushanta Karmakar, Intrusion Detection Systems using Linear Discriminant Analysis and Logistic Regression, in INDICON. 2015, IEEE.

[7] Thulasy Ramiah Pillai, A.A., Sellappan Palaniappan,Hafiz Muhammad Imran, Predictive Modeling for Intrusions in Communication Systems using GARMA and ARMA models. 2015.

[8] Cho, J., et al., Dynamic learning model update of hybrid-classifiers for intrusion detection. The Journal of Supercomputing, 2011. 64(2): p. 522-526.

[9] WU YANG, W.W., LIN GUO, LE-JUN ZHANG, AN EFFICIENT INTRUSION DETECTION MODEL BASED ON FAST INDUCTIVE LEARNING, in Proceedings of the Sixth International Conference on Machine Learning and Cybernetics. 2007: Hong Kong. p.3249-3254

[10] Tylman, W., Anomaly-Based Intrusion Detection Using Bayesian Networks. 2008: p. 211-218.

[11] Tylman, W., Misuse-Based Intrusion Detection Using Bayesian Networks. 2008: p. 203-210.

[12] Liyuan Xiao, Y.C., and Carl K. Chang, Bayesian Model Averaging of Bayesian Network Classifiers for Intrusion Detection.pdf, in IEEE 38th Annual International Computers, Software and Applications Conference Workshops. 2014. p. 128-133.

[13] Ployphan Sornsuwit, S.J., Intrusion Detection Model Based on Ensemble Learning for U2R and R2L Attacks. 2015. P.354-359

[14] Sumaiya Thaseen, C.A.K., An Analysis of Supervised Tree Based Classifiers for Intrusion Detection System, in International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME). 2013, IEEE. p. 294-299.

[15] Sumouli Choudhury, A.B., Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection, in International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM). 2015. p. 89-95.

[16] Zhang Hongmei, SVM Ensemble Intrusion Detection Model Based on Rough Set Feature Reduct, Chinese Control and Decision Conference. 2009. p.5604-5608.

[17] YANG Ya-Hui, HUANG Hai-Zhen, SHEN Qing-Ni, WU Zhong-Hai, ZHANG Ying, Research on Intrusion Detection Based on Incremental GHSOM, Chinese journal of computers, 2014.37(5):p.1216-1224.

[18] WANG Chao, XIN Yang, Application of C5.0 Algorithm to Network Intrusion Detection, The 13th Annual Meeting of China Association for Science and Technology. 2011.p.178-184.

[19] Classifier-trees-RandomTree on weka source code analysis [Online] Available:http://blog.csdn.net/roger__wong/article/details/39272625(may 15, 2016)