International Journal of Research in Engineering, Science and Management
Volume-2, Issue-11, November-2019
www.ijresm.com | ISSN (Online): 2581-5792

Consolidation of C 5.0, Random Forest and Random Tree Crafts for Intrusion Detection System

Meghana Solanki[1], Trupti Phutane[2]
[1,2]*Assistant Professor, Department of Computer Engineering, D. Y. Patil College of Engineering, Pune, India*

***Abstract*: Our day to day activities e.g. ecommerce are influenced by Internet. The hazards from hackers have also expanded. According to view point of many researchers intrusion detection systems are the base of defense. There are many commercially available intrusion detection systems are principally signature-based. Known attacks are detected by these systems. These systems frequently renovate signature as well as rules. Unknown attacks are not detected by these systems. The use of anomaly base intrusion detection systems are the best solution. They are highly potent in recognizing not only known but also unknown attacks. Recognition of high false alarm rate is main problem with anomaly based intrusion detection systems. In our proposed system, we contribute explanation to increase attack detection rate while reducing high false alarm rate by combining various data mining techniques such as C 5.0, Random Forest and Random Tree.**

***Keywords*: C5.0, Data Mining, Intrusion Detection System, Random Forest, Random Tree.**

1. Introduction

Any collection of actions that try to negotiate the virtue, affection, opportunity of a system is called an intrusion. Computer attacks are detected by Intrusion Detection system (IDS). It examines different log as well as data records. Host-based attacks [19-21] as well as network-based attacks [22-24]; these are the two category of attack. In Host based attacks a specific machine is targeted by attacker. They try to seek approach to either restricted services or resources. Host-based detection makes the use of not only audit process but also data for attack detection. Network-based attacker willfully occupy network resources as well as services to restrict legitimate users from access various network services. This can be completed by sending heavy amounts of network transit as well as by taking benefit of known faults. It also overloads network hosts. In Network-based attack detection, intrusion detection made by analyzing network traffic. Basically there are two verity of anomaly detection system i.e. Misuse base and anomaly base [24, 25]. First one relies on blueprint and second one relies on learning as well as training the normal behavior of a system. In case of first type, human expertise is required for generation of specification or rules. These systems are simply expansion of misuse base IDS systems. Snort as well as Bro generally used rule-based network intrusion detection systems. They contain rules which are written manually for identification of known attacks. Detection of virus and probing attack is done by adding manually virus signatures and permission for accessing services or hosts. Anomaly base intrusion detection systems are very useful in detecting known as well as unknown attacks. Detection of high false alarm rate is the problem with anomaly base intrusion detection systems. In our proposed system, we tried to solve this problem by combining three data mining techniques C 5.0, Random Forest and Random Tree.

2. Literature survey

In this paper [1], an author provided detail extensive analysis of anomaly detection techniques by using machine learning as well as demographic modes. In this paper [2], an author gave review of not only numeric but also symbolic data for anomaly detection. In these papers [3] [4], an author conferred an extensive audit of detection techniques using neural networks as well as statistical method. In these papers [12] [5], an author presented survey for cyber-intrusion detection by using various anomaly detection techniques. In these papers [6]-[9], an author have given various anomaly detection systems such as NIDES, ALAD, PHAD, and SPADE generate statistical models for normal network traffic as well as set up alarms when there is a deviation from the normal model. In these papers [10], [11], an author presented exhaustive reviews of several anomaly detection methods. In this paper [12], an author analyzed not only IP traffic but also presented major techniques but also problems for application detection. In this paper [13], an author presented survey for network anomaly detection methods as well as techniques. In this paper [14], an author presented not only review of flow-based intrusion detection but also explained concepts of flow and classified attacks. They also analyzed in detail techniques used for detection of scans, worms, Botnets and DoS attacks. In this paper [15], an author gave overview of Intrusion detection techniques and methods for mobile ad-hoc networks (MANET) and wireless sensor networks (WSN). In this paper [16], an author given exhaustive survey of techniques for detecting DoS as well as distributed DoS attack. In this paper [17], an author gave overview of

computational intelligence methods for intrusion detection using various techniques such as swarm intelligence, soft computing, evolutionary computation, fuzzy systems, artificial neural networks, and artificial immune systems. In this paper [18], an author explained an application layer IDS using sequence learning for detection of anomalies.

3. Data Mining and NSL-KDD Data Set

A. *Data Mining*

Data mining also known as Knowledge Discovery in Databases – KDD. It is used for data processing using sophisticated data search capabilities as well as statistical algorithms to discover designs and interrelationships in large preexisting databases. It is new way to discover new meaning in data. It is nontrivial eradication of tacit, previously unknown, and probably useful information from data. It is also called as data discovery or knowledge discovery. In case of data mining techniques, data from many different perspectives or dimensions or angles are allowed to analyze from user. They categorize them as well as summarize the identified relationships. It is the process of extracting not only relations but also patterns among various fields in databases. The main three important steps in data mining are Extract, Transform, and Load (ETL). One of the most generally used data mining responsibility is classification. Classification model is built using training data set, which describes the data classes or concepts. The classification model is used to do prediction of objects class. The classification model is built training data sets as well as test data set. The build model can be presented in the form of decision trees or mathematical formulae or rule or neural networks. This paper is proposed on three base classification techniques mainly C5.0, Random Tree and Random Forest.

B. *C5.0 Tree*

C5.0 method is expanded by Quinlan which is mainly based on C4.5 algorithm. It contains not only latest automation but also the most important application is “boosting” technology. The most prominent methods in intrusion detection system are random forest method as well as random tree. Why do we prefer C5.0 method in the proposed system? The C5.0 method has not only good disclosure veracity but also a short disclosure era. We also find that C5.0 method conduct work with both continuous as well as categorical components. Furthermore, they are impressive against redundant along with correlated variables. They are pivotal to handle the 42 features of NSLKDD dataset.

C. *Random Forests*

Leo Breiman and Adele Cutler created Random Forests algorithm. It is an assemblage learning method for classification as well as regression. It erects a number of decision trees (CART) at training time. Also they are not altered by each other. While allegation it bulks all predication made by all decision trees. It is mostly used for the reasoning of complex data structures which contains large column data with small to moderate data sets.

D. *Random Tree*

It formulates tree using K randomly selected attributes at every node beyond pruning. It estimates of class contingencies depend on a hold-out set.

E. *NSL-KDD data set*

The NSL-KDD data set contains 42 attributes. It is used in this factual study. This data set is an enhancement over KDD’99 data set. It removed duplicate instances from KDD’99 data set to get rid of biased classification results. This data set has number of versions available, out of which 20% of the training data is used. NSL-KDD data set consists of four major attacks categories and they are as follows;

- DOS (Denial of Service): This attack can chill the server operation and activity. It acquires all resources so that the server cannot any afford any service, commonly using flooding based blueprints.
- PROBE: This attack is used during a formation stage for other attacks in order to gain antique information such as enabled ports and services. It also gains Internet address information.
- U2R (User to Root): This attack executes a peculiar operation in order to peek into a system hole/leak such as Buffer Overflow.
- R2L (Remote to User): The attack is designed to take benefits of safety information of users or configuration such as SQL Injection. Number of records in training data set as well as testing data set is displayed in Table 1.

Table 1
Number of Instances

Type	Training Data Set	Test Data Set
DOS	486268	348942
PROBE	5219	5375
U2R	69	339
R2L	2234	27291
NORMAL	223441	71682
Total	717231	453629

4. Experiments and Results

We perform experimentation using NSL-KDD data set with 717231 records and test record set using WEKA 3.8.3. For our experiments Intel Core i5, 3.4 GHz processor with 8 GB RAM as a hardware and Windows 8 64 bit, WEKA 3.8.3 as a software. We use two specifications, attack detection rate as well as false attack detection rate for evaluating the performance of our approach. Attack detection rate is determined as total attack detected using combination of data mining algorithms split by total number of attacks in test data set. False attack detection rate is determined as total no of attack

instance detected as normal instance using combination of data mining algorithms spilt by total number of group wise attack instance in test data set. We can evaluate our approach with the help of these two parameters. These parameters predict us what proportion of intrusion is detected by our approach. It also predicts how many incorrect classifications it can make. The results gained from proposed system are demonstrated below. The confusion matrix provided by WEKA is used to obtain results. Table 2 shows details of correctly as well as incorrectly categorized instances. Table 3 shows category wise attack detected.

Table 2
Number of Classified Instances

Classifiers	Classified Instances	
	Correctly	Incorrectly
C 5.0	93.52	6.48
Random Forest	93.56	6.44
Random Tree	91.42	6.58

Table 3
Division Wise Attack Detected

Classifiers	DOS		PROBE		U2R	
	Correct	False	Correct	False	Correct	False
C 5.0	334799	7276	4253	2135	8	332
Random Forest	**335043**	**7032**	**4360**	1028	3	337
Random Tree	327941	24134	3973	2425	36	314

From table 3 it is clear that C5.0 implements superior in detecting Normal division. Random Forest implements superior in DOS and PROBE type attack division. Random Tree implements superior in U2R and R2L type attack division.

Table 4
Division Wise Attacks Detected Using Mixed Algorithm

Combination	C 5.0 and Random Tree		Random Forest and Random Tree		C 5.0 and Random Forest	
Attack Category	Correct	False	Correct	False	Correct	False
U2R	39	311	38	312	10	330
R2L	2862	25549	2855	25556	2396	26015
PROBE	4419	969	4437	951	4436	952
DOS	335009	6966	335072	6903	335113	6962
NORMAL	71468	345	60792	1021	71415	398

From Table 4 it is clear that the union of C 5.0 as well as Random Tree implements superior than any other combination except in PROBE type attack division. From Table 4 we can say that the combination implements superior than any individual algorithm in all types of attack division by using attack detection rate as well as false attack detection rate. Table 5 shows analogy of our proposed system with entries of NSL KDD tournament. We got superior results in DOS, R2L and Normal division. Also we are marginally trailing in case of U2R as well as Probe division.

Table 5
Analogy with Entries of NSL-KDD Tournament

Combination of Classifiers	Attack Category				
	DOS	PROBE	U2R	R2L	Normal
Entries	334337	**4582**	**41**	2471	71373
C 5.0 and Random Tree	**335309**	4419	39	**2862**	**71468**
Random Forest and Random Tree	334082	4437	38	2855	71233
C 5.0 and Random Forest	335113	4436	10	2396	71233

5. Conclusion

This paper demonstrates idea of mixing of data mining algorithms for enhancing attack detection rate as well as reducing false attack detection rate. We explained the results of mixing C 5.0 with Random Tree, C 5.0 with Random Forest, and Random Forest with Random Tree classifiers. Also results are outlined in Table 4. All these experiments were implemented using not only NSL KDD data set with full attributes but also WEKA 3.8.3 tools. After auditing results using performance parameters we wind up that mixing C 5.0 with Random Tree enhances performance of intrusion detection by making use of both the parameters. The performance was improved due to random tree which implements superior in U2R and R2L type of attack. C5.0 implements superior in detecting Normal division. Also it is marginally trailing in case of U2R as well as Probe division compared to others classifier. Thus by taking benefits of both classifier we can achieve better attack detection rate. Our approach achieved better results in DOS, R2L and Normal attack division.

References

[1] V Hodge, and J Austin, "A survey of outlier detection methodologies." Artificial Intelligence Review 22, 2, 2004, 85-126.
[2] M Agyemang, K. Barker, and R Alhajj "A comprehensive survey of numeric and symbolic outlier mining techniques", Intelligent Data Analysis 10, 6, 521-538, 2006.
[3] M. Markou, and S. Singh "Novelty detection: a review-part 1: statistical approaches", Signal Processing 83, 12, 2481-2497, 2003.
[4] A. Patcha and J. Park "An overview of anomaly detection techniques", Existing solutions and latest technological trends. Comput. Networks 51, 12, 3448-3470, 2007.
[5] P. Rousseeuw, and L. Seroy "Robust regression and outlier detection" John Wiley & Sons, Inc., New York, NY, USA, 1987.
[6] H. Javits and A. Valdes. "The NIDES statistical component" Description and justification. Technical report, SRI International, Computer Science Laboratory, 1993.
[7] M. Mahoney "Network Traffic Anomaly Detection Based on Packet Bytes" Proc. ACMSAC 2003.
[8] M. Mahoney, P. K. Chan. "Learning Nonstationary Models of Normal Network Traffic for Detecting Novel Attacks", Proc. SIGKDD 2002, 376-385, 2002.
[9] J. Hoagland, "SPADE", Silican Defense, http://www.silicondefense.com/software/spice, 2000.
[10] V. Chandola, A. Banerjee, and V. Kumar: "Anomaly Detection: A Survey", ACM Computing Surveys, vol. 41, no. 3, 15:1–15:58, September 2009.
[11] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez, and E. Vazquez "Anomaly-based network intrusion detection" Techniques, systems and challenges, Computers & Security, vol. 28, no. 1-2, 18–28, 2009.
[12] C. Callado, C. Kamienski, G. Szabo, B. Gero, J. Kelner, S. Fernandes, and D. Sadok, "A Survey on Internet Traffic Identification", IEEE Commun. Surveys Tutorials, vol. 11, no. 3, 37–52, 2009.

[13] W. Zhang, Q. Yang, and Y. Geng "A Survey of Anomaly Detection Methods in Networks," in Proc. International Symposium on Computer Network and Multimedia Technology, 1–3, January 2009.
[14] G. Sperotto, R. Schaffrath, R. Sadre, C. Morariu, A. Pras, and B. Stiller: "An Overview of IP Flow-Based Intrusion Detection", IEEE Commun. Surveys Tutorials, vol. 12, no. 3, 343–356, 2010.
[15] I. Sun, Y. Osborne, Xiao, and S. Guizani "Intrusion detection techniques in mobile ad hoc and wireless sensor networks", IEEE Wireless Commun., vol. 14, no. 5, 56–63, October 2007.
[16] T. Peng, C. Leckie, and K. Ramamohanarao "Survey of network-based defense mechanisms countering the DoS and DDoS problems", ACM Computing Surveys, vol. 39, no. 1, 1–42, April 2007.
[17] S. X. Wu and W. Banzhaf "The use of computational intelligence in intrusion detection systems: A review", Applied Soft Computing, vol. 10, no. 1, 1–35, January 2010.
[18] Y. Dong, S. Hsu, S. Rajput, and B. Wu, "Experimental Analysis of Application Level Intrusion Detection Algorithms", International J. Security and Networks, vol. 5, no. 2/3, 198–205, 2010.
[19] S. Axelsson: "Research in intrusion detection systems: a survey." Technical Report TR 98-17. Chalmers University of Technology, Goteborg, Sweden (1999).
[20] D. Anderson, T. Frivold, A. Valdes "Next-generation intrusion detection expert system (NIDES)" a summary. Technical Report SRI-CSL-95-07. Computer Science Laboratory, SRI International, Menlo Park, CA (May 1995).
[21] Freeman, S., Bivens, A., Branch, J., Szymanski, B.: Host-based intrusion detection using user signatures. In: Proceedings of the Research Conference. RPI, Troy, NY, 2002.
[22] K. Ilgun, R.A Kemmerer, P.A. Porras, "State transition analysis: A rule-based intrusion detection approach." IEEE Trans. Software Eng. 21(3), (1995), 181–199.
[23] D. Marchette, "A statistical method for profiling network traffic", In: Proceedings of the First USENIX Workshop on Intrusion Detection and Network Monitoring, Santa Clara, CA (1999) 119–128.
[24] S. Mc Canne, C. Leres, Jacobson "Libpcap", available via anonymous ftp at ftp://ftp.ee.lbl.gov/ (1989).
[25] S. Mukkamala, G Janoski, A Sung "Intrusion detection: support vector machines and neural networks", In: Proceedings of the IEEE International Joint Conference on Neural Networks (ANNIE), 1702– 1707. St. Louis, MO (2002).