Indian Journal of Science and Technology, Vol 9(33), DOI: 10.17485/ijst/2016/v9i33/97037, September 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Hybrid Scheme based on Big Data Analytics using Intrusion Detection System

Shaik Akbar*, T. Srinivasa Rao and Mohammed Ali Hussain

Department of CSE, Andhra Loyola Institute of Engineering College, Vijayawada - 520008, Andhra Pradesh, India; dr.shaikakbar@gmail.com, Srinivas123fast@gmail.com, alihussain.phd@gmail.com

Abstract

Objective: Network security plays a key role for many organizations. Host based and network based Intrusion Detection Systems are available in the market depending upon the detection technology used by them. The objective of this research paper is maintaining security across the heterogeneous data from homogeneous sources and co-relating the heterogeneous data from different sources using hybrid strategy. **Methods/Statistical Analysis:** A real time detection Intrusion Prevention Systems (IPS), prevents security intrusions by gathering and composing with technologies. **Findings:** Heterogeneous data from different sources has been collected from KDD Cup Dataset and segregated into learning phase and detection phase. In the learning phase, known attacks will be identified. Similarly detection phase also will consider the same. **Applications/Improvements:** The proposed system specifies a set of rules and high DoS, R2L, U2R, Probe. One may attempt to get good results by improving the efficiency and reducing the complexity present in the model. In future several reduction techniques may be studied to get more features.

Keywords: Big-Data, Host Based, IDS, Network Based, Security

1. Introduction

As computerized systems are increasing day by day in the aspects of finances, industry, medicine etc., it is very heptic task to maintain the cyber security. Intrusion Detection is one of the most important considerations of cyber security. Even for forensic purposes, Intrusion Detection will be used in-order to indentify successful breaches even after they have occurred.

To detect attacks or abnormal conditions Intrusion Detection can be very helpful in cyber area. The traditional IDS evaluate the network traffic using a network layer of OSI reference model. Similar to these traditional IDS, one more IDS which is host based IDS monitor's cyber attacks by monitoring a host system logs, system process and files. Analysis of Big-data will be considered by IDS. Big-data is always a challenging issue in streaming the large quantity of data using computing technologies. Hence, IDS can be used for deep packet inspection[1]. Now a days, many organizations are facing to have an incredible amount of log data[2] and this issue can be efficiently managed by Intrusion Detection System.

This paper mainly solves the two issues of the following

- Maintaining security across the heterogeneous data from homogeneous sources.
- Co-relating the heterogeneous data from different sources using hybrid strategy.

1.1 IDS and Big Data Survey

Here mainly discuss about the selection of available data sets and attributes of it. There are 3V's that can be mainly defined in Big data[3,4] volume, velocity and variety. Volume defines the quantity of data and it can be a Big data defiance. When huge amounts of data pose defiances to processing with different computing technologies. Velocity generalizes the speed at which data is processed and there can be a Big data issue when the rate of data is moving costly with computing technologies. Variety explains about the complexity of the data and this is also a Big data defiance when

*Author for correspondence

the data contains difficult problems such as high spatiality, data from heterogeneous origins or data having plenty of different data structures. In addition to this, two more V's velocity and value can be considered[5]. Velocity generalizes the accuracy of the data and can possess data correctness problems as squeak or missing values. Value defines Big-data such that if concerned data does not produce significant value. Perhaps, being solved all the five V's, Big data issues by Intrusion Detection, mainly discusses on data redundancy and set of different patterns[6]. For example if a user produces five to forty Mega bytes of data in an eight hour slot and it takes plenty of hours to analyze a single hours bundle of data. They further explains that clustering, filtering and feature selection on the data is vital if real time detection is expected which can improve detection correctness. As on today, need the improvements in security systems, to detect security intrusions and these improvements can be considered[7] by Intrusion Prevention Systems (IPS) which prevent security intrusions by gathering and composing with technologies. Actually, an IPS will work close to real time detection.

In the previous era, security monitoring was executed by system administrators verifying the log files of their servers. Once IDS came into existence, a separate monitoring device performs all the validations at the network or host level. Even though, IDS is having[8] the Big-data problem associated with large volume of data collected from intrusion detection dataset[9,10]. Selecting the datasets is always a challenging issue. Many of the researchers admit that the underlying datasets[11] themselves have inherent flaws[12–14]. A brief analysis is also given on feature selection and its application to Intrusion Detection datasets. This can be important especially whenever using datasets from heterogeneous sources. In-order to address Big-data challenges from Intrusion Detection, always feature selection is an important technique. This technique improves classification accuracy by removing noise. A few numbers of features will improve classification processing times from an efficiency standpoint. However, it will take certain amount of computation time for feature selection. Applying feature selection is not an easy and it may not able to generate feature sets in close to real time. In-terms of classification accuracy, feature selection technique on KDD-Cup99[15–17] achieves the best overall feature selection results as compared to eleven other techniques. Since the feature selection takes certain amount of computation time, and this time can be reduced by half when using SVM[18–20] and C4.5 on KDD datasets.

2. Proposed Architecture

Heterogeneous data from different sources has been collected from KDD Cup Dataset and segregated into learning phase and detection phase. In the learning phase always identifies known attacks. In-order to identify these known attacks to select the best attributes from the KDD Cup Dataset. Similarly detection phase also will consider the same. The Figure 1 output of feature selection will be applied as an input to the preprocessing and here the data will be cycled in various stages. To identify the known attacks i.e., DoS, Probe, R2L,U2R and the attacks which are not identified are treated as detection phase. In this detection phase using enhanced C4.5 and enhanced Genetic Algorithm were used. These two techniques were storing the data in-terms of database independently.

These two databases were hybrid together forming an integrated large database. The output of this large database will be applied as an input to Big-data using data streams one pass constraint technique using this technique we will generate the classifying the dataset corresponding to this dataset we can find the relative type of attack which are DoS, Probe, R2L and U2R. Finally, we can conclude that our hybrid scheme will increase the detection rate and also identifies the type of the attack belongs to.

3. Experimental Setup

The scope of our experiment has been to focus on generating classifiers on rules for 25 attack types belonging to four different categories, and for creating a rule that can classify all of these connections with a minimal false positive rate.

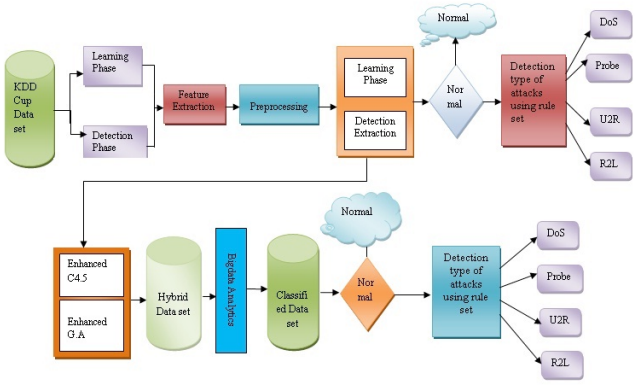

Figure 1. Akbar: Proposed hybrid intrusion detection system architecture.

Table 1. Shows the detection and false positive rate for hybrid scheme

Sl. No	Attack Category	Detection Rate (%) (C4.5 + G.A)	False Positive (%) (C4.5 + G.A)
1	DoS	94.80	0.043
2	Probe	96.42	0.082
3	U2R	95.52	0.063
4	R2L	90.87	0.045
Average Success Rate		94.40	0.058

The assessments demonstrate the power and capability of the planned method for performing extremely well with 94.80% detection for DoS attacks, 90.87% detection for R2L attacks, 95.52% detection for U2R attacks and 96.42% detection for Probe attacks with an overall FPR of 0. 058%. The test similarity of this technique has testified to its utility and importance. Our algorithm has resulted in agreeable choices of FAR and the major development in D.R for all kinds of assaults with an overall D.R of 94.40% which have resulted with very huge dataset and suboptimal ingredient IDSs as shown in Table 1.

4. Conclusions

The main contribution of this work is to develop dynamic IDS by using machine learning techniques is sequence to better identify anomalies and to reduce false positive rate. The proposed Genetic Algorithm presents the Intrusion Detection System for detecting DoS, R2L, U2R, Probe from KDDCUP99 Dataset. The outputs of the experiments are satisfactory with an average success rate of 92.595% and the overall results of the technique implemented are good.

The proposed system is useful in different areas with more flexibility and good attack taxonomy. With the growing complexity of the intrusions and rapid changes, the Intrusion Detection System should compete with the thread space. The correlation techniques recognize the network connections and Enhanced Genetic Algorithm detects the Intrusion. The experimental results specify that Enhanced Rule based G.A gives better accuracy than C4.5 algorithm for DoS, Probe, U2R and R2L classes. New methods should be studied and their effectiveness should be calculated as I.D replicas.

5. Future Work

With the growing incidents of pretend assaults, structure efficient I.D replicas with good rightness and real-time performance are essential. Additional Data Mining methods should hence be checked for a more fruitful attribute extraction. The proposed system specifies a set of rules and high DoS, R2L, U2R, Probe. One may attempt to get good results by improving the efficiency and reducing the Complexity present in the model. In future several reduction techniques may be studied to get more features.

6. Acknowledgements

Fr. Dr. A. Francies Xavier SJ, Secretary and Director of Andhra Loyola Institute of Engineering and Technology, Vijayawada Supports us in making this paper a very highly achievable hybrid system.

Fr. M. L. Thomas SJ, Assistant Director of Andhra Loyola Institute of Engineering and Technology, Vijayawada stood background to our paper and supports in all the aspects in making this executable one.

7. References

1. Center for Strategic and International Studies. The economic impact of cybercrime and cyber espionage. Technical report. McAfee. Available from: http://www.mcafee.com/us/resources/reports/rp-economic-impact-cybercrime.pdf
2. A case study in security big data analysis. Available from: http://www.darkreading.com/analytics/security-mointoring.a-case-study-in-security-big-data-analysis/d/d-id/1137299
3. Denning DE. An intrusion-detection model. Soft Eng IEEE Trans SE. 1987; 13(2):222–32.
4. Frank J. Artificial intelligence and intrusion detection: Current and future directions. Proceedings of the 17th National Computer Security Conference; Baltimore, MD, USA. 1994. p. 1–12.
5. Group BDW big data analytics for security intelligence. Available from: http://downlods.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_intelligence.pdf
6. Information assurance solutions group, defense in depth. Technical report; 2015. National Security Agency. Available from: http://www.nsa.gov/ia/_files/support/defenseindepth.pdf
7. Julisch K, Dacier M. Mining intrusion detection alarms for actionable knowledge. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining. ACM; Edmonton, Alberta, Canada. 2002. p. 366–75.
8. Data management: Controlling data volume, velocity and variety. Technical Report 949; META Group (now Gartner). Available from: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
9. Nassr M, Bouna AB, Mallui Q. Secure outsourcing of network flow data analysis. IEEE International Congress on Big Data; Santa Clara, CA, USA. 2013; 2(3):431–2.
10. Ponemon Institute LLC. Cost of cyber crime study: United States. Technical Report. Ponemon Institute. Available from: http://www.ponemon.org/local/upload/file/2012_US_Cost_of_Cyber_Crime_Study_FINAL6\%20.pdf
11. Roesch M. Snort: Lightweight intrusion detection for networks. LISA. USENIX; Seattle, WA, USA. 1999. p. 229–38.
12. Sourcefire, Snort, Home Page. Available from: http;//www.snort.org/
13. Verizon RISK Team, data breach investigations report. Technical Report. Verizon. Available from: www.verizonenterprise.com/verizon-insights-lab/dbir/
14. Zikopoulos P, Parasuraman K, Deutsch T, Giles J, Corrigan D. Harness the power of big data the IBM big data platform. New York, NY: McGraw Hill Professional; 2012.
15. Arora SK, Vijan S, GabaGS. Detection and analysis of black hole attack using IDS. Indian Journal of Science and Technology. 2016 May; 9(20). DOI: 10.17485/ijst/2016/v9i20/85588.
16. Srikanth BVS, Reddy VK. Efficiency of stream processing engines for processing BIGDATA Streams. Indian Journal of Science and Technology. 2016 Apr; 9(14). DOI: 10.17485/ijst/2016/v9i14/84797.
17. Kyoo-sung N, Doo-sik L. Bigdata platform design and implementation model. Indian Journal of Science and Technology. 2015 Aug; 8(18). DOI: 10.17485/ijst/2015/v8i18/75864.
18. Renjit JA, Shunmuganathan KL. Network based anomaly intrusion detection system using SVM. Indian Journal of Science and Technology. 2011 Sep; 4(9). DOI: 10.17485/ijst/2011/v4i9/30239.
19. Azad C, Jha VK. Data mining based hybrid intrusion detection system. Indian Journal of Science and Technology. 2014 Jun; 7(6):781–9.
20. Mourougan S, Aramudhan M. Hybrid evolutionary algorithm based intrusion detection system for denial of service attacks. 2015 Dec; 8(35). DOI: 10.17485/ijst/2015/v8i35/86652.