International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP—71-77

Security Analytic on Big Data: A Classification Technique to Detect Intrusion

[1]Nilamadhab Mishra,[2]Sarojananda Mishra
[1]CSE , Ph.D. Scholar, BPUT, Odisha, India
[2]IGIT, Sarang , Dhenkanal ,Odisha, India

Abstract: *To maintain the stability of any network, Intrusion Detection plays a vital role. Big Data Security Analytics is a process of searching, analyzing and recognizing the required patterns from large amount of data. The major requirements for any intrusion detection system are speed, accuracy and less memory. Although various intrusion detection methods are available, they work at some points while lack in the others. It presents a comprehensive survey of the technologies which are used for detecting intrusions. It analyzes the advantage and disadvantage of each technology and the literature works that utilizes these technologies. Challenges faced on the current IDS and the requirements for IDS in the current network scenario are discussed. The research framework is proposed and a discussion of the different technologies that can be used for improving the efficiency of the IDS is provided.*

Keywords*: Big data, Data Mining, Intrusion Detection System (IDS), Machine Learning.*

I. Introduction

Today we are living in an era of digital world. With the rapid increase in digitization the amount of Structured, semi structured and unstructured data being generated and stored is exploding. In Big data the information comes from multiple, heterogeneous, autonomous sources with complex relationship and continuously growing. Every day, 2.5 quintillion bytes of data are created daily and 90 percent data in the world today were produced within past two years [4].for example Flicker, a public picture sharing site, where in an average 1.8 million photos per day are receive from February to march 2012[13]. Currently Big Data processing depends upon parallel programming models like Map Reduce, as well as providing computing platform of Big Data services. Thus making big data mining or knowledge discovery of large datasets a difficult process. Data mining algorithms need to scan through the training data for obtaining the statistics for solving or optimizing model parameter. The most fundamental challenge for big data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. [2]

There are different types of data such as relational, structural, textual, semi structured, graph data, streaming data etc can be included in big data. Big Data Analytics is defined as the process of analyzing and understanding the characteristics of massive size datasets by extracting useful geometric and statistical patterns. Ideally these three characteristics of a dataset increase the complexity of the data and thus make the current techniques and technologies stop functioning as expected within a given processing time. Many applications suffer from the Big Data problem, including network traffic risk analysis, geospatial classification and business forecasting. Network intrusion detection and prediction are time sensitive applications and they require highly efficient Big Data techniques and technologies to tackle the problem on the fly. The new technologies can help conduct Big Data analytics on various applications. [5] The problems and challenges associated with the integration of modern networking technologies and machine learning techniques for solving Big Data classification problem for network intrusion prediction.

II. Literature review

In the year 2015, authors reviewed studies on the data analytics from the traditional data analysis to the recent big data analysis. From the system perspective, the KDD process is used as the framework for these studies and is summarized into three parts: input, analysis, and output. From the perspective of big data analytics framework and platform, the discussions are focused on the performance-oriented and results-oriented issues. From the perspective of data mining problem, this paper gives a brief introduction to the data and big data mining algorithms which consist of clustering, classification, and frequent patterns mining technologies. To better understand the changes brought about by the big data, this paper is focused on the data analysis of KDD from the platform/framework to data mining. The open issues on computation, quality of end result, security, and privacy are then discussed to explain which open issues we may face. [1]

Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and bio-medical sciences. This author presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. Authors analyze the challenging issues in the data-driven model and also in the Big Data revolution. [4] This paper suggested an integration of modern technologies, Hadoop Distributed File Systems and Cloud Technologies, with the latest representation-learning technique and support vector machine to predict network intrusions through Big Data classification strategy. Additionally it suggested adopting machine lifelong learning framework for solving the problems associated with the continuity parameter. It also discussed the problems and challenges that the Big Data classification system for network intrusion prediction have to experience during the Big Data analytics. [5] Useful data can be retrieved from this large datasets with the aid of big data mining [10]. Here the data which are handled is big data, hence the term big data mining. Usually, data mining is the technique of analyzing data from different prospects and summarizing these data into interesting, understandable and useful models.

Thus a large attempt to exploit these huge parallel processing architectures was initiated. Big [8] data mining must deal with heterogeneity, extreme scale, velocity, privacy, accuracy, trust, and inter activeness that existing mining techniques and algorithms are incapable of the need for designing and implementing very-large-scale parallel machine learning and data mining algorithms (ML-DM)has remarkably increased, which accompanies the emergence of powerful parallel and very-large-scale data processing platforms, e.g., Hadoop Map Reduce. NIMBLE [11] is a portable infrastructure that has been specifically designed to enable rapid implementation of parallel MLDM algorithms, running on top of Hadoop. Apache's Mahout [12] is a library of machine learning and data mining implementations. The library is also implemented on top of Hadoop using the Map Reduce programming model. Some important components of the library can run stand-alone. The main drawbacks of Mahout are that its learning cycle is too long and its lack of user-friendly interaction support.

BC-PDM (Big Cloud-Parallel Data Mining) [14], as a cloud-based data mining platform, also based on Hadoop, provides access to large telecom data and business solutions for telecom operators; it supports parallel ETL process (extract, transform, and load), data mining, social network analysis, and text mining. BC-PDM tried to overcome the problem of single function of other approaches and to be more applicable for Business Intelligence. PEGASUS (Pet-scale Graph Mining System) and Graph both implement graph mining algorithms using parallel computing and they both run on top of Hadoop. Graph Lab is a graph-based, scalable framework, on which sever all graph-based machine learning and data mining algorithms are implemented. [7]

III. Big data

Big data refers to large data sets that are challenging to store, search, share, visualize and analyze. It is high volume, high velocity, high variety information assets that demand cost effective, innovative forms of information processing for enhance insight and decision making.[3] Big data is coined to address massive volumes of data sets usually huge, sparse, incomplete, uncertain, complex or dynamic, which are mainly coming from multiple and autonomous sources. The 3Vs that define big data are Volume, Velocity and Variety [9].

Volume

Volume means vast amount of data generated in every second [10]. It is a scale characteristic. The data is in rest state. Machine generated data are examples for these characteristics. Nowadays data volume is increasing exponentially.

Velocity

The second generated characteristics of big data are velocity or speed. Velocity is the speed at which data generated. The streaming data may not be massive and its state is in motion. It should have high speed data. Example is data created through social media. The data is begin generated fast and need to be processed fast. Online Data Analytics includes these types of big data. E-Promotions and health care monitoring are examples. In e-promotion, based on our current location and our purchase history, what we like will send promotions right now for store next to us. In Healthcare monitoring, sensors monitoring our activities and body. Any abnormal measurements require immediate reaction can be immediately identified through this.

Variety

Variety is another important characteristic of big data. Various data formats, types, and structures can be referred here. The type of data may include different verities such as Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…It also includes s static data and streaming data . A single application can be generating by collecting many types of data. To extract the knowledge all these types of data need to be linked together.

Now two more V's also contributed to big data. They are veracity and value of data [6].

Veracity

Veracity means data in doubt. The uncertainty of data can be found due to the inconsistency and incompleteness. The messiness of data (Abbreviation, colloquial speech etc) may result the veracity.

Value

Value gives importance to the profit gained by organizations who invest in Big Data technologies.

IV. Big data mining

Useful data can be retrieved from this large datasets with the aid of big data mining [10]. Here the data which are handled is big data, hence the term big data mining. Usually, data mining is the technique of analyzing data from different prospects and summarizing these data into interesting, understandable and useful models. For better decision making, the large repositories of data collected from different resources require a proper mechanism for extracting knowledge from the databases. Since big data scales far beyond the capacity of single PC, cluster computers, which have high computing powers and rely on parallel programming paradigms, are used. Thus a large attempt to exploit these huge parallel processing architectures was initiated. Big [8] data mining must deal with heterogeneity, extreme scale, velocity, privacy, accuracy, trust, and inter activeness that existing mining techniques and algorithms are incapable of the need for designing and implementing very-large-scale parallel machine learning and data mining algorithms (ML-DM)has remarkably increased, which accompanies the emergence of powerful parallel and very-large-scale data processing platforms, e.g., Hadoop Map Reduce. NIMBLE [11] is a portable infrastructure that has been specifically designed to enable rapid implementation of parallel MLDM algorithms, running on top of Hadoop.

Apache's Mahout [12] is a library of machine learning and data mining implementations. The library is also implemented on top of Hadoop using the Map Reduce programming model. Some important components of the library can run stand-alone. The main drawbacks of Mahout are that its learning cycle is too long and its lack of user-friendly interaction support. Besides, it does not implement all the needed data mining and machine learning algorithms. BC-PDM (Big Cloud-Parallel Data Mining) [14], as a cloud-based data mining platform, also based on Hadoop, provides access to large telecom data and business solutions for telecom operators; it supports parallel ETL process (extract, transform, and load), data mining, social network analysis, and text mining. BC-PDM tried to overcome the problem of single function of other approaches and to be more applicable for Business Intelligence. PEGASUS (Pet-scale Graph Mining System) and Graph both implement graph mining algorithms using parallel computing and they both run on top of Hadoop. Graph Lab is a graph-based, scalable framework, on which sever all graph-based machine learning and data mining algorithms are implemented. Distinctive algorithms used in data mining are as follows:[7]

Classification trees:

A famous data-mining system that is used to categorize a needy categorical variable based on size of one or many predictor variables. The outcome is a tree with links and nodes between the nodes that can be interpret to form if-then rules.

Logistic regression:

A algebraic technique that is a modification of standard regression but enlarges the idea to deal with sorting. It constructs a formula that predicts the possibility of the occurrence as a role of the independent variables.

Neural Networks:

A software algorithm that is molded after the matching architecture of animal minds. The network includes of output nodes, hidden layers and input nodes. Each unit is allocated a weight. Data is specified to the input node, and by a method of trial and error, the algorithm correct the weights until it reaches a definite stopping criteria. Some groups have likened this to a black– box system.

Clustering techniques like K-nearest neighbors:

A procedure that identifies class of related records. The K-nearest neighbor technique evaluates the distances between the points and record in the historical data. It then allocates this record to the set of its nearest neighbor in a data group.

Machine Learning

The traditional Machine Learning (ML) techniques have been developed and used for extracting useful information from the data through training and validation using labeled datasets. Three major problems that make the ML techniques unsuitable for solving Big data classification problems are: (1) An ML technique that is trained on a particular labeled datasets or data domain may not be suitable for another dataset or data domain that the classification may not be robust over different datasets or data domains; (2) An ML technique is in general trained using a certain number of class types and hence a large varieties of class types found in a dynamically growing dataset will lead to inaccurate classification results; and (3) An ML technique is developed based on a single learning task, and thus they are not suitable for today's multiple learning tasks and knowledge transfer requirements of Big Data analytics.

V. Intrusion detection system

Increase in the amount of data transfer in networked environments, especially the Internet has led to an increase in the potential threats. With the cost of processing getting decreased from time to time, adversaries are gaining more prominence and are exploiting the system vulnerabilities further. This has led to the development of mechanisms to counter the attacks, called the Intrusion Detection Systems (IDS). The major functionality of IDS is to monitor and analyze traffic, identifying abnormal activities and assessing the severity of the situation and raising alarm. Figure 1 shows the architecture of typical IDS.

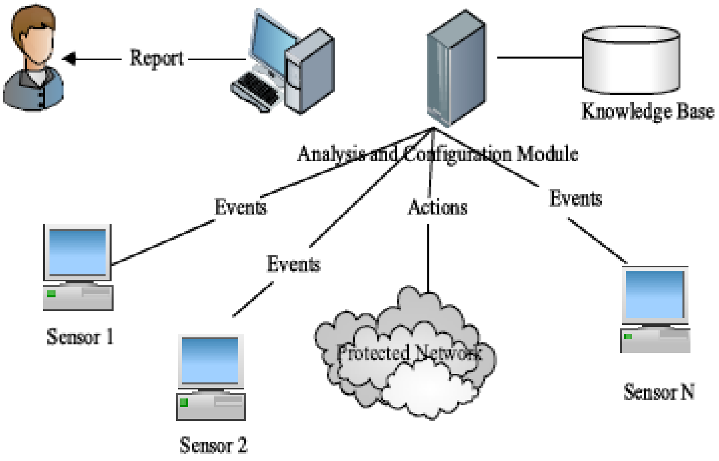

Fig 1: Architecture of typical IDS.

The major components of an IDS are the nodes/sensors on which the events take place. The events can correspond to normal activity or malicious activities. These events are recorded by the analysis & configuration module, which uses the knowledge base for categorizing the traffic as normal or anomalous. Reports are generated based on the analysis and is presented to the user for further analysis. The performance variables that play a vital role in determining the efficiency of the system are the detection rate (DR) and the false alarm rate (FAR).[15]

Intrusion Detection using Big Data Analytics:

The recent years have seen tremendous researches in big data, which is due to the increase in the information flow in the network. The large amount of traffic (log data) generated from networks (volume) and the speed at which the data is generated (velocity) is sufficient to justify the usage of Big Data technologies for the process of intrusion detection. This is a new direction, and research literatures in this area are very less. Hadoop is used as the standard environment for developing Big Data applications. The availability of various algorithms in the Hadoop environment has proved to be a major positive aspect, which attracts researches and researchers towards this technology. The availability of parallelization options in this area is an added advantage. An adaptive detection approach for detecting anomalies using big data is presented and leverages the parallelization facilities available in the MapReduce to perform effective classification of network data.[15]

Table.1. Summary of IDS/IPS Techniques[15]

IDS/IPS Technique	Characteristics/ Advantage	Limitation/Challenges

Signature based detection	• Identifies intrusion by matching captured patterns with preconfigured knowledge base. • High detection accuracy for previously known attacks. • Low computational cost.	• Cannot detect new or variant of known attacks. • High false alarm rate for unknown attacks.
Anomaly detection	• Uses statistical test on collected behavior to identify intrusion. • Can lower the false alarm rate for unknown attacks.	• More time is required to identify attacks. • Detection accuracy is based on amount of collected behavior or features.
ANN based IDS	• Classifies unstructured network packet efficiently. • Multiple hidden layers in ANN increase efficiency of classification.	• Requires more time and more samples training phase. • Has lesser flexibility.
Fuzzy Logic based IDS	• Used for quantitative features. • Provides better flexibility to some uncertain problems.	• Detection accuracy is lower than ANN.
Association rules based IDS	• Used to detect known attack signature or relevant attacks in misuse detection.	• It cannot detect totally unknown attacks. • It requires more number of database scans to generate rules. • Used only for misuse detection.
SVM based IDS	• It can correctly intrusions, if limited sample data are given. • Can handle massive number of features.	• It can classify only discrete features. So, preprocessing of those features is required.
GA based IDS	• It is used to select best features for detection. • Has better efficiency.	• It is complex method. • Used in specific manner rather than general.
Hybrid techniques	• It is an efficient approach to classify rules accurately.	• Computational cost is high.

VI. Objectives

- To classify the network traffic data for intrusion prediction.
- To increase classification accuracy.
- To provide suitable computational complexity for Big Data Analytics.

VII.Methodologies

Different classification Techniques:

- Classification by decision tree induction
- Bayesian Classification
- Rule based Methods
- Memory based reasoning
- Neural Networks
- Support Vector Machine

Proposed Technique: Improvised SVM approach

VIII. Conclusion

In real-world applications managing and mining Big Data is Challenging task. The latest representation-learning technique and support vector machine to predict network intrusions through Big Data classification strategy. Additionally it suggested adopting machine learning framework for solving the problems associated with the continuity parameter. It also discussed the problems and challenges that the Big Data classification system for network intrusion prediction have to experience during the Big Data analytics. Research on Big Data techniques and technologies evolving and at the same time new problems and challenges are emerging, hence the hope is to develop better and better techniques and technologies towards finding solutions for Big Data classification problem.

References

[1]. Chun Wei Tsai1, Chin Feng Lai2, Han Chieh Chao1,3,4 and Athanasios V. Vasilakos5, Big data analytics: a survey, Tsai et al. Journal of Big Data (2015) 2:21 ,Springer Open Journal , DOI 10.1186/s40537-015-0030-3.

[2]. Manisha V.Kharat1,Dhiraj V. Bhise2 ,Data Mining With Big Data: A Servey Paper,International Journal of Advance Research in Computer Science and Management Studies, ISSN: 2321-7782 (Online),Volume 3, Issue 7, July 2015.

[3]. Neelamani Samal, and Nilamadhab Mishra, "Big Data Processing: Big Challenges and Opportunities." *Journal of Computer Sciences and Applications*, vol. 3, no. 6 (2015): 177-180. doi: 10.12691/jcsa-3-6-13.

[4]. Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu, and Wei Ding" Data Mining with Big Data" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014 .

[5]. Shan Suthaharan, University of North Carolina at Greensboro,Greensboro, NC 27402, USA, Performance Evaluation Review, Vol. 41, No. 4, March 2014.

[6]. Anuradha, G., "Suggested techniques for clustering and mining of data streams", Published in: Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on Date of Conference:4-5 April 2014.

[7]. Vitthal Yenkar, 2Prof.Mahip Bartere," Review on Data Mining with Big Data", Vitthal Yenkar et al, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.4, April- 2014, pg. 97-102 ,ISSN2320–088X.

[8]. B R Prakash1*, Dr. M. Hanumanthappa 2," Issues and Challenges in the Era of Big Data "ISSN 2278-6856, Volume 3, Issue 4 July-August 2014 Mining .

[9]. SHERIN A1, Dr S UMA2, SARANYA K3, SARANYA VANI M4" SURVEY ON BIG DATA MINING PLATFORMS, ALGORITHMS AND CHALLENGES" Sherin A et al. / International Journal of Computer Science & Engineering Technology (IJCSET), ISSN : 2229-3345 ,Vol. 5 No. 09 Sep 2014 .

[10]. SMITHA T, V. Suresh Kumar, "Application of Big Data in Data Mining" ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 7, July 2013 .

[11]. NewVantage Partners: Big Data Executive Survey (2013) http://newvantage.com/wpcontent/ Uploads/2013/02/ NVP-Big- Data-Survey- 2013-Summary-Report.pdf .

[12]. Xin, R.S., Rosen, J., Zaharia, M., Franklin, M., Shenker, S., Stoica, I.: Shark: SQL and Rich Analytics at Scale. In: ACM SIGMOD Conference (accepted, 2013) .

[13]. F. Michel, "How Many Photos Are Uploaded to Flicker Every Day and Month?" http://www.flickr.com/photos/franckmichel/6855169886/, 2012.

[14]. Agrawal, D., Bernstein, P., Bertino, E., et al.: Challenges and Opportunities With big data Community White Paper Developed by Leading Researchers Across the United States (2012), http://cra.org/ccc/docs/init/bigdatawhitepaper. Pdf .

[15]. S.J. Sathish Aaron Joseph, R. Balasubramanian ,"A Comprehensive Survey of Technologies for Building a Hybrid High Performance Intrusion Detection System",*International Journal of Computer Applications,Volume 113– No.15,March 2015.*