Hadoop based Parallel Machine Learning Algorithms for Intrusion Detection System

Malathi Eswaran, P. Balasubramanie, M. Jotheeswari

Abstract: ***Web use and digitized information are getting expanded each day. The measure of information created is likewise getting expanded. On the opposite side, the security assaults cause numerous security dangers in the system, sites and Internet. Interruption discovery in a fast system is extremely a hard undertaking. The Hadoop Implementation is utilized to address the previously mentioned test that is distinguishing interruption in a major information condition at constant. To characterize the strange bundle stream, AI methodologies are used. Innocent Bayes does grouping by a vector of highlight esteems produced using some limited set. Choice Tree is another Machine Learning classifier which is likewise an administered learning model. Choice tree is the stream diagram like tree structure. J48 and Naïve Bayes Algorithm are actualized in Hadoop MapReduce Framework for parallel preparing by utilizing the KDDCup Data Corrected Benchmark dataset records. The outcome acquired is 89.9% True Positive rate and 0.04% False Positive rate for Naive Bayes Algorithm and 98.06% True Positive rate and 0.001% False Positive rate for Decision Tree Algorithm.***

Keywords -Big data, Decision Tree, Hadoop Framework, Intrusion Detection, Machine Learning, Naïve Bayes.

I. INTRODUCTION

As of late, the mechanical advancement has enhanced the web speed from terabytes to petabytes. Individuals from different fields, for example, specialized and non-specialized people are getting benefits by utilizing Internet Services and its applications. All the business fields are mechanized and in this way creating computerized information over the system. Thus, tremendous volume of information is produced from different fields, for example, sensor information, video information, sound information and versatile information and so on. Intrusion is any unlawful PC movement that gets access for data in a system or an association. Moreover, it is the course of action of exercises that tries to deal the dependability, security and openness of a benefit in a framework. The interruption counteractive action strategies, for example, encryption, validation, get to control, secure steering are putting forth security against a few assaults, for example, Denial of administration (Dos) assault, client to root assault (U2R), remote to nearby assault (R2L) and testing assaults. Assaults are of two kinds in particular dynamic and latent.

Revised Manuscript Received on November 06, 2019.

Malathi Eswaran, Department Computer Technology – PG, Kongu Engineering College, Erode

P. Balasubramanie, Department of Computer Science & Engineering Kongu Engineering College, Erode

M. Jotheeswari, Department Computer Technology – PG, Kongu Engineering College, Erode

Interruption Detection System (IDS) is for the most part worked for dynamic sort of assaults. For recognition instrument, the IDS can be abnormality based, abuse based, have based, arrange based and cross breed based. Inconsistency based identification is amazing for most recent assaults that are obscure and the assault is experienced out of the blue [1]. Peculiarity based Intrusion Identification System could be established on accurate estimations in which the framework traffic is gotten and depiction is maintained. The depiction is stream based. Deviation from the specific limit of inconsistency score is recognized as an interruption. One of the real strategies utilized in peculiarity based discovery is Machine Learning (ML) [2]. The Machine Learning approach utilized in this paper is Naïve Bayes Classification and Decision Tree strategies. Innocent Bayes Classifier is one of the directed learning model [3]. Regulated learning considers the issues of evaluating a model from information tests with name data. Credulous Bayes classifier accept that the estimations of explicit highlights are autonomous of alternate highlights in a class. Choice Tree is another ML classifier which is in like manner a coordinated learning model [4]. It is a stream outline like tree structure where each inward focus exhibits a test on a property, each branch tends to a delayed consequence of the test and each leaf holds a class name. The most vital focus point in a tree is root focus point. The Hadoop MapReduce Model is the parallel preparing system for Big Data Environments. This structure is actualized to parallelize the Machine Learning Algorithm. The Naïve Bayes and the Decision tree classifiers are executed in MapReduce that parallelize the bigger datasets like KDDCup Benchmark datasets and arrange the anomalous parcels stream with higher location rate.

II. RELATED WORK

The significant bind in the Big Data condition is the treatment of vast volume of information and the conventional consecutive programming worldview couldn't deal with various assortments of information. Distinguishing interruption in such enormous information conditions with the assistance of machine learning calculations are examined in this segment. Idris et al. proposed a method for the identification of different sorts of assaults, for example, DoS assault, wormhole assault. The standard based strategy is utilized, that depends upon known degree augmentation appear by delineating the power ruin of the message transmission rule [5]. This method deal with the message as suspicious if its transmitted energy gets crashed its sender's topographical position. Erfan A.Shams and Ahmet Rizaner proposed the Bayesian Classification real machine that is utilized to see the interruption in the framework [6]. Their guideline factor is to apprehend the pack flooding that results in Denial of Service (DoS).The proposed mannequin makes use of a direct showcase that keeps up the a variety of clients depiction. For keeping up

such profiles they applying back Bayesian arrangement. This Bayesian Classification Algorithm is considered as the recognition calculation that identifies flooding assaults. Natesan P and Rajalakshmi R proposed a parallel figuring model and a nature enlivened element choice strategy, a Hadoop Based Parallel Binary Bat Algorithm machine is proposed for profitable issue assurance and canny portrayal to get multiplied distinguishing proof rate [7]. The MapReduce programming model of Hadoop improvements computational multifaceted nature, the Parallel Binary Bat be counted improves the features decision and parallel Naïve Bayes Algorithm offers canny gathering.

Ioannis Krontiris et al proposed framework based intrusion disclosure structure (NIDS) for the canny sensor-breathed life into contraption [8]. Structure primarily based IDS watched and made an appraisal on each pushing toward heap of the shape site visitors and noticed the take a look at that passed off in the system. This can be finished on framework contraptions, for instance, switch, switch, server, entry, etc., They proposed a NIDS embedded in a splendid sensor-propelled contraption under an organization masterminded building (SOA) approach which can work self-governing as a peculiarity based NIDS, or facilitated direct in a Distributed Intrusion Detection System (DIDS).This joins the upsides of the shrewd sensor approach and the ensuing offering of the NIDS usefulness as an administration with the SOA use to accomplish their combination with different DIDS segments. It lessen the enormous volume of the executives undertakings intrinsic to this kind of system administrations, just as encouraging the plan of DIDS whose overseeing unpredictability could be confined inside all around characterized edges. Sun Mei-Feng and Chen Jing-Tao proposed the idea of exploring the traffic qualities reasonable for the on-line traffic grouping [9]. The traffic order serves to recognizing the application related with TCP stream [10]. It takes the entire statistics length, which is sent with the aid of the amigo before it received the first ACK bundle, as the properties. They consider two properties ACK-Len solid quality and ACK-Len ba. ACK-Len is utilized to address the complete information length from A to B before the first ACK pack showed up. Then again, implied ACK-Len ba. To check the adequacy of ACK-Len strong quality and ACK-Len ba, they used choice tree C4.5 figuring as classifier to see four sorts of utilization, for example, WWW, FTP, EMAIL and P2P.

III. PROBLEM DESCRIPTION

A tremendous section of the proposed Intrusion Detection System for different sorts of ambushes was not fit perceive darken attacks. Some of the Machine Learning algorithms was unable to produce better detection rate for various attacks. Also the larger volume of data and the space required for handling those big data could not be addressed by the previous mechanisms which are sequential programming paradigm. The main objective is to parallelize the processing for such larger volume of data using the Hadoop Environment and hence the space requirements are handled efficiently by the MapReduce Framework.

IV. PROPOSED SYSTEM

Hadoop is the framework for handling large amount of data and run applications on large clusters which is built on commodity hardware. Machine Learning algorithms are highly utilized in the field of Spam Filtering and the Intrusion Detection. Hadoop MapReduce Framework along with the Machine Learning Algorithms gives better interpretation towards various types of data and can able to handle enormous amount of data. The two ML algorithms in MapReduce programming framework and Dataset used for evaluation as shown below. 1. Naive Bayes Classification, 2.Decision Tree Algorithm, 3.Dataset Description, 4. Real Time Datasets using network tools, 5. Multinode Cluster, 6. Evaluation Parameters.

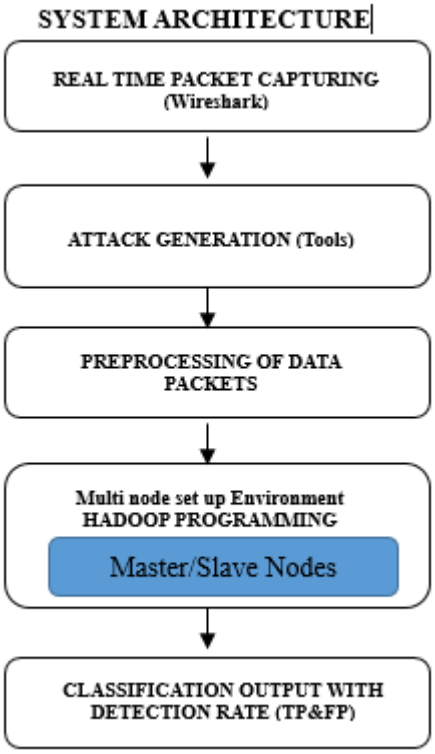

Fig.1 System architecture.

A. Native Bayes Classification

The algorithm pseudocode has two separate mapper and reducer functions. First the mapper function executed completely then the reducer function will execute and finally produce the result. Based on the size of input file the Hadoop Mapreduce Framework will separate the whole file into different chunks for efficient processing and all the chunks are executed separately and finally the result is the combination of the output from all the chunks. The both mapper and reducer functions will take the input in the form of key/value pairs. The pseudo code is given in Fig.2 and Fig.3.

B. Decision Tree

J48 is an enlargement of ID3 decision tree. The additional features of J48 are speaking to missing characteristics, decision trees pruning, consistent attribute regard ranges, finding of guidelines, etc. The J48 Decision tree classifier seeks after going with fundamental computation. In order to portray something else, it first needs to choose a choice tree subject to the quality estimations of the accessible

arranging information. Along these lines, at whatever point it encounters a ton of things (getting ready set) it perceives the quality that isolates the diverse precedents for the most part evidently. The part which can uncover to us most about the information occasions with the target that we can portray them the best is said to have the most basic data gain. Before long, among the potential estimations of the fragment, if there is any a helper for which there is no helplessness, that is, for which the information cases falling inside its group have a similar force for the objective variable, by then inferring that branch and assign to the objective respect that acquired.

Pseudo code 1: Naïve Bayes- Map (key, value)

Input : key<input filename> ; values<Instances>
Output: key'<classes> ; value'<conditional probability>
/* for attributes A1,A2,.....A_{dp} and class label C1,C2,.....

C_l of dataset D_p*/

FOR j ← 1 to l
FOR i ← 1 to dp
$p(C_j|A_i) \leftarrow p(A_i|C_j) * p(C_j)$;

END FOR
END FOR
key' <classes>;
value' <conditional probability for all classes>
emit <key',value'>

Fig.2 Mapper Function for the Naive Bayes Algorithm

Pseudo code 2: Naïve Bayes- Reduce (key, value)

Input: key<classes>; values<Conditional Probability for various classes>
Output: key'<classes>; value'<highest conditional probability>
/* highest conditional probability for given conditional variable X*/
FOR j ← 1 to l
  FOR i ← 1 to m
    calculate $p(C_j|X_i)$ from all map functions;
  END FOR
    label $(C_j) \leftarrow$ highest $[p(C_j|X_i)]$
END FOR
  key' <classes>;
  value' <highest conditional probability >
  emit <key',value'>

Fig.3 Reducer Function for the Naive Bayes Algorithm

C. Dataset description

The KDDCup99.Corrected.Data Benchmark dataset which is the subset of DARPA dataset is utilized for assessment. KDDCup99 getting ready dataset is around four giga bytes of compacted twofold TCP dump data from seven weeks of system traffic, managed into around 5,000,000 association records each with around 100 bytes. The two weeks of test information have around 2,000,000 alliance records. Each KDDCup99 planning association record contains 41 solidifies and is isolated as either traditional or a trap, with unequivocally one express assault type.

D. Real Time Datasets

The real time datasets are generated by real time packet capturing tools such as ettercap, snort and wireshark.

(i) Ettercap Tool

It is free open source compose security instrument for man-in-the-inside ambushes on LAN. It will by and large be utilized for PC driving force show assessment instrument and for security reviewing. It continues running on various UNIX like working structures including Linux, Mac OS, Solaris and Microsoft windows. It is fit for blocking traffic on a framework divide, getting passwords and moreover coordinating unique tuning in against different typical shows. It offers four strategies for action, for instance, IP based, MAC based, ARP based and Public ARP based.

(ii) Snort Tool

Snort is very powerful tool for packet capturing and can able to generate DDOS attacks. Snort can be run in following four modes. Sniffer mode, parcel lumberjack mode, IDS mode and IPS mode.

(iii) Wireshark

Wireshark is a system bundle analyzer, referred to already as Ethereal. It enable us to inspect the system traffic, streaming into and out of windows or Unix machine. Wireshark is for the most part utilized for investigating system issues.

E. Multinode Cluster

The proposed system is implemented in Hadoop multinode cluster. The Multinode cluster gives the memory efficient parallel processing and it can handle the large volume of real time data. The multinode cluster separates the large input file and give that to several slave nodes that is multiple clusters and provides results with higher accuracy and efficiency compared to single node environment. Multinode with 2 nodes and 3 nodes are used to test the real time datasets which also gives higher accuracy and efficiency.

F. Evaluation Parameters

The evaluation parameters used to evaluate the proposed work are Detection Rate, False positive Rate and Accuracy.

$$Detection\ Rate = \frac{TP}{TP + FN}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP}$$

TN- True Negative, TP- True Positive, FP- False Positive, FN- False Negative.

IV. RESULTS AND DISCUSSION

The proposed framework is assessed by utilizing both the

Benchmark KDDCup99.Corrected.Data datasets and Real time datasets in Hadoop singlenode and multinode condition. The model got outcomes as far as obvious positive rate and false positive rate. The Naïve Bayes calculation gives 89.9% genuine positive rate and 0.004% false positive rate. The J48 Decision tree gives 98.06% genuine positive rate and 0.001% false positive rate. The graphical variety of the two calculations with Hadoop singlenode and multinode as shown in Fig.4 and Fig.5

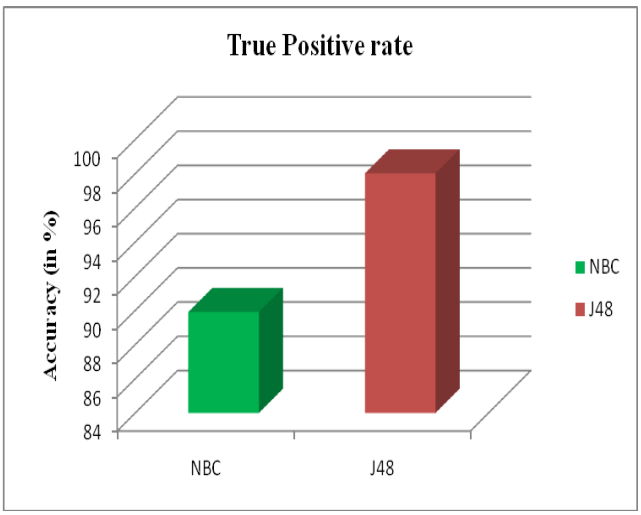

Fig.4 True positive variation between Naïve Bayes and Decision Tree.

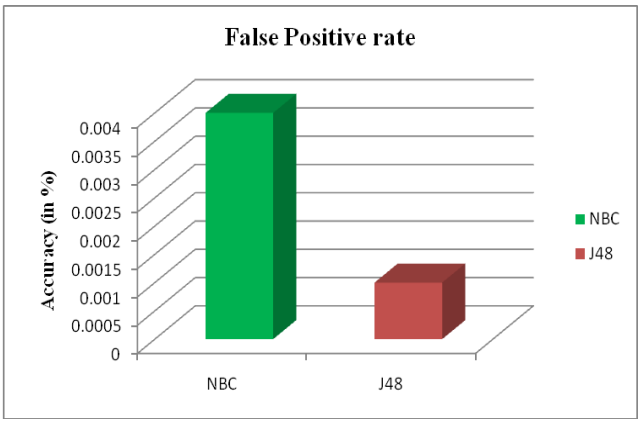

Fig.5 False positive variation between Naïve Bayes and Decision Tree.

Efficiency comparison between hadoop singlenode and multinode setup in terms of execution time as shown in Fig.6 and Fig.7

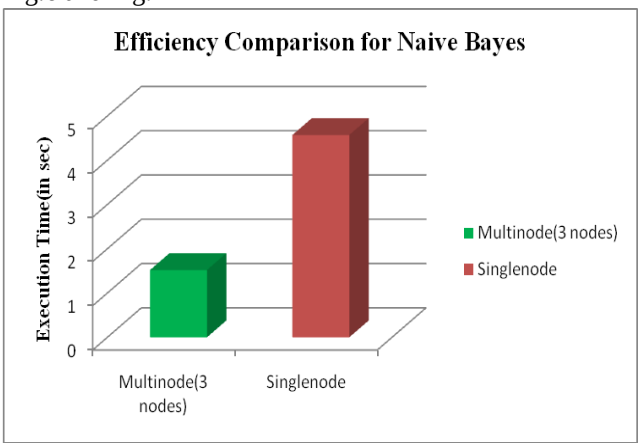

Fig.6 Efficiency comparison between hadoop single node and multinode for Naïve Bayes Algorithm.

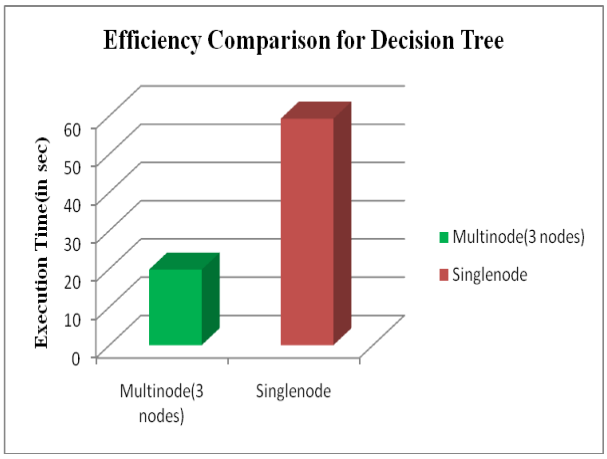

Fig.6 Efficiency comparison between hadoop single node and multinode for Decision Tree Algorithm.

V. CONCLUSION

This work exhibited two Machine Learning calculations Naive Bayes and Decision Tree for the system interruption identification framework. This calculation utilizes the Hadoop Map Reduce system for parallel preparing of the bigger datasets. The most broadly utilized datasets, for example, KDDCup99 and KDD Cup Corrected datasets and the ongoing datasets caught from system parcels (bundle catching instruments, for example, Wire shark, Etter cap, Snort) are utilized for assessment and testing of the framework. The order indicated precision as far as True Positives and False Positives which is nearer to the exactness. The Naïve Bayes calculation gives 89.9% genuine positive rate and 0.004% false positive rate. The J48 Decision tree gives 98.06% genuine positive rate and 0.001% false positive rate. This is nearer to the precision. To build the precision and effectiveness, Machine Learning Algorithms will be executed in Hadoop Multi node Setup Environment. In addition the Hadoop Framework will be arranged into multi node with the goal that the vast volume of information from the constant system traffic can be taken care of proficiently. In future, to accomplish the continuous productivity, the Spark device will be utilized over the Hadoop Ecosystem.

REFERENCES

1. M. Usha and P. Kavitha, "Anomaly based intrusion detection for 802.11 networks with optimal features using SVM classifier", Wireless Networks, vol. 23, no. 8, 2016, pp. 2431-2446.
2. T. Abbes, A. Bouhoula and M. Rusinowitch, "Efficient decision tree for protocol analysis in intrusion detection", International Journal of Security and Networks, vol. 5, no. 4, 2010, p. 220.
3. A. Kaur, S. Pal and A. Singh, "Hybridization of K-Means and Firefly Algorithm for intrusion detection system", International Journal of System Assurance Engineering and Management, vol. 9, no. 4, 2017, pp. 901- 910.
4. W. Yu and H. Lee, "An Incremental-Learning Method for Supervised Anomaly Detection by Cascading Service Classifier and ITI Decision Tree Methods", Intelligence and Security Informatics, 2009, pp. 155-160.

5. S. Lim and Y. Choi, "Malicious Node Detection Using a Dual Threshold in Wireless Sensor Networks", Journal of Sensor and Actuator Networks, vol. 2, no. 1, 2013, pp. 70-84.
6. E. Shams and A. Rizaner, "A novel support vector machine based intrusion detection system for mobile ad hoc networks", Wireless Networks, vol. 24, no. 5, 2017, pp. 1821-1829.
7. P. Natesan, R. Rajalaxmi, G. Gowrison and P. Balasubramanie, "Hadoop Based Parallel Binary Bat Algorithm for Network Intrusion Detection", International Journal of Parallel Programming, vol. 45, no. 5, 2016, pp. 1194-1213.
8. I. Krontiris, T. Giannetsos and T. Dimitriou, "LIDeA", Proceedings of the 4th international conference on Security and privacy in communication netowrks - SecureComm '08, 2008.
9. M. Sun and J. Chen, "Research of the traffic characteristics for the real time online traffic classification", The Journal of China Universities of Posts and Telecommunications, vol. 18, no. 3, 2011, pp. 92-98.
10. M. Zhao, B. Jia, J. Wang, M. Wu and H. Yu, "Performance Optimization on Dynamic Adaptive Streaming over HTTP in Multi-User MIMO LTE Networks", IEEE Transactions on Mobile Computing, vol. 17, no. 12, 2018, pp. 2853-2867.

AUTHORS PROFILE

Malathi Eswaran received B.E, M.E degrees in 2009, 2011 respectively from Avinashilingam University, Coimbatore and Kongu Engineering College (Autonomous), Erode, Affiliated to Anna University, Chennai. She has 6 years of teaching experience in Engineering Colleges. Her research interests are Data Mining, Machine Learning and Big Data.

P. Balasubramanie MSc., MPhil., M.Tech., Ph.D is currently working as a Professor in the department of computer Science & Engineering, Kongu Engineering College, Perundurai, Tamilnadu, India. He is one of the approved supervisors of Anna University, Chennai and guided 26 Ph.D scholars. Currently he is guiding 9 Ph.D scholars. He has published 218 articles in International/National journals.

He has authored/co-authored 11 books with the reputed publishers. He has completed one AICTE RPS as a Principal investigator and currently he is working as a principal investigator in a MRP of UGC. He has received Rs. 13 Lakhs of grant from various funding Agencies like AICTE, CSIR, NBHM, DRDO, INSA and so on and organized 21 STTP/SDP/Seminar/Workshops for the benefit of Faculty members and Research scholars. He has received 17 awards so far from various agencies. His area of interest includes Data mining, Networking, cloud computing and Optimization algorithms Learning and Big Data.