IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Network Anomaly Detection on Fast Streaming Data Using Spark

D. Priya

Student, Computer Science and Engineering, Bharathidasan University, Trichy, India

Abstract: Intrusion detection is a continuous process and depending on the size of the network and the number of transmissions being carried out in the network, the number of packets to be analyzed varies considerably. Hence there is no specific or defined data size associated with it, but the Velocity component of Big Data plays a vital role here. The packets being transferred tends to be fast, hence a mechanism to provide analysis in real time becomes mandatory. This paper presents a technique to predict intrusions faster and with higher accuracy. It uses a Random Forest based classifier implemented on Hadoop platform using Spark. Spark, being a stream processing framework exhibits effective results in real-time.

Keywords: Intrusion Detection; Networks; Hadoop; Spark; Random Forest.

I. INTRODUCTION

Cyber security has become a critical aspect due to the increase in the use of computers in all industries such as finances, medicine, industry etc. A major requirement in cyber security is Intrusion Detection. This acts as a major contributor in identifying attacks or malicious behaviour. Increase in efficiency of this system will lead to a better and more secure online environment. Online transactions have become one of the day to day activities due to the advent of ecommerce. Most of the ecommerce organizations prefer to store sensitive information, so that it becomes easier for the user to perform transactions [1]. It becomes mandatory for the organizations to store this information in a secure manner. Any information available in digital form is prone to hacks. Due to the importance associated with the contained information, the process of detection and prevention should be synonymous to the attack. Existing intrusion detection and prevention systems tend to provide a buffer of a few transactions before identifying the intrusion, which tends to be costly. Further, network data tends to be huge, hence analyzing the entire data is both time and processor consuming [2]. Hence it becomes mandatory for the systems to either use Big Data techniques or reduce the available data to chunks that can be handled by the existing processing architecture. This approach presents an architecture to identify intrusions in real time, combined with the flexibility of using even very huge amounts of data. The ever increasing data plays a huge threat to this model. With the increase in usage of digital media, the data generated by them have also increased to a large extent. Hence faster processing with traditional approach is not feasible. It requires specific Big Data based approaches in order to effectively process them and provide results. A data is classified as Big Data if it effectively satisfies any one of the requirements of Big Data, namely; Volume, Velocity and Variety. Real time data generated satisfies the constraint of Big Data, hence it is classified as Big Data.

II. RELATED WORKS

A profiling based intrusion detection system that uses network traffic to identify intrusions is presented in [3]. Network packets tend to occur in large numbers. Analyzing each packet for intrusion is not feasible. This method is based on a reduction strategy that eliminates probably legitimate packets and passes only a few packets for processing. Alpha and beta profiling are used to reduce the number of data for comparison. Feature based reduction is also performed to reduce the number of comparisons further. A statistical rule based intrusion detection mechanism is presented in [8]. This is a genetic algorithm based technique that is designed to evolve a set of simple interval based rules. Genetic algorithm is modified such that the rule set is maintained small. A similar method concentrating on DDoS attacks is presented in [9]. A similar technique using GA that speeds up the detection mechanism is presented in [15]. Algorithms to speedup pattern matching have also been analyzed and documented in [7], which provides efficient techniques to improve the process of pattern matching. These techniques were effectively utilized in the process of intrusion detection. Similar rule based techniques were in prevalence. Apriori algorithm based intrusion detection is presented in [10].

A collaborative method for intrusion detection in mobile networks is proposed in [4]. This method mostly concentrated on stealth attacks which cannot be detected by any existing intrusion detection mechanisms. A multi-level intrusion detection mechanism is presented in [5]. This method uses a coarse grained and a fine grained mechanism to identify intrusions. Examining each packet for intrusion is very tedious and hence not feasible. The coarse grained detection mechanism is activated initially and checks for intrusions. A packet, if identified as probable intrusion, is taken to the fine grained control for extended evaluation. A priority based intrusion detection mechanism is presented in [12]. This method also presents

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

post correlation techniques to analyze the results obtained. Neighbor based intrusion detection methods [13] that operate by ranking the neighbors have been widely used in clustered environments. Similar to statistical models, machine learning models are also on the raise, due to the ever changing nature of the intrusions. An analysis of such mechanisms is presented in [14]. A combined SVM and PSO based intrusion detection method that also uses dimensionality reduction is presented in [11]. Increase in the number of processor cores and reduction in cost of processors has lead to an increased use of parallelization techniques. A survey of intrusion detection techniques on GPUs is presented in [6]. A parallel intrusion detection method that uses GPGPUs to perform faster and energy efficient intrusion detection is presented in [6].

III. OUR APPROACH

The input data is preprocessed to convert it to a format accepted by the classifier. The operations carried out in this stage is minimal, because random forest classifiers handles missing data and imbalanced data well. Hence there is no necessity to eliminate or handle missing data. The data is then segregated to training and test data, to aid in the cross validation testing that is to be carried out on the classifier. The training data is then passed to the random forest classifier for building the classifier model.

The random forest classifier is an ensemble model that uses several decision trees on subsets of the data to build decision rules. Subset creation is carried out in such a manner that each of the subsets contain at least 66% of the original data. This is to make sure that all the classes contained in the original data have representatives in the subset. This would ensure that all the classes are considered prior to the splitting section in the decision trees.

The next phase of this process is the actual creation of the decision trees. Data subsets are passed to the decision trees. Each decision tree identifies a subset of m predictor variables from the M total predictor variables (m<M). The best predictor variable is identified from this set and a binary split is performed on it. This marks the beginning of the tree creation. The process of predictor variable identification and splitting is carried out for all the available predictors and the decision tree is constructed. Pruning is not performed on the decision tree in order to retain all the data during the rule aggregation step.

There exist three different methods to select the value of m. The Random splitter selection method, where m=1, the Breiman's bagger method, where m=M and the Random Forest method where m<<M. Brieman suggested three possible ways of selecting the values of m, namely $^1/_2\sqrt{m}$, $\sqrt{m}$ or $2\sqrt{m}$.

Each of the rules are obtained by training the decision trees on a part of the data, hence the rules returned by each of the decision trees differ. The final classifier model is built by combining all these rules. It has been observed that though each of the decision trees provided weak rules, a combination of these rules exhibits a strong classifier with improved reliability and accuracy.

IV. RESULTS AND DISCUSSION

Experiments were carried out on a Hadoop 2 using Spark. The machine learning library (MLlib) was used to create Random Forest for the process of Intrusion detection. Experiments were carried out on the KDD Cup 99 dataset and the results were recorded.

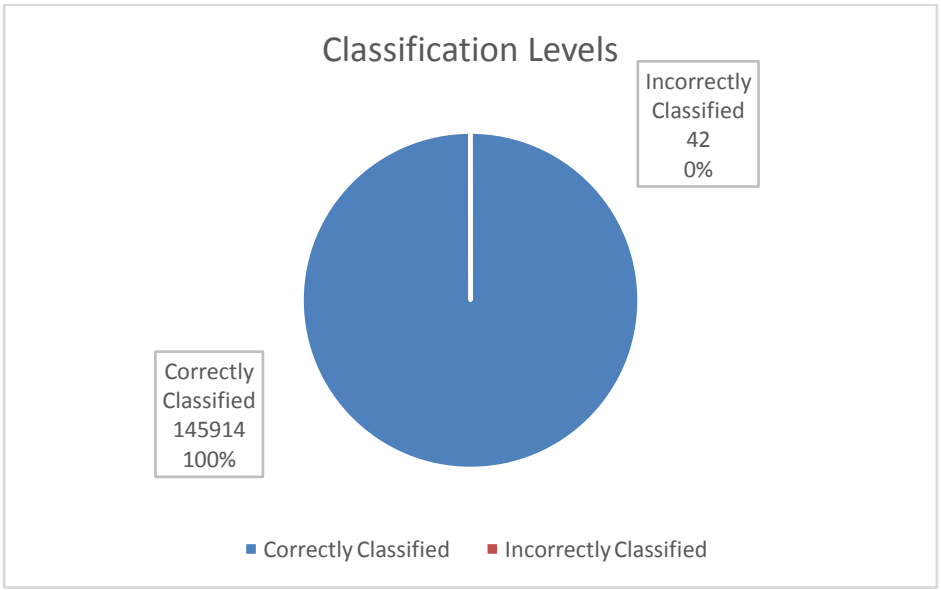

Fig.1 Classification Levels

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

The classification levels obtained from Random Forest are presented in Figure. It could be observed that the correct classification levels outperform the incorrect classification levels. The correct classification percentage was observed to be 99.97% and the incorrect classification levels were rated only to 0.03%. Hence it could be concluded that Random Forest operates efficiently on the network data to produce high quality results.

Naïve Bayes is one of the most popular classifier algorithms existing in the stream processing architecture. Hence an accuracy comparison between the Random Forest and Naïve Bayes is performed (Figure). It could be observed that, Naïve Bayes, being a probabilistic technique exhibits lower accuracy of ~50%, while Random Forest exhibits an accuracy of 99%. This exhibits the efficiency of tree based algorithms when operated on huge streaming data.

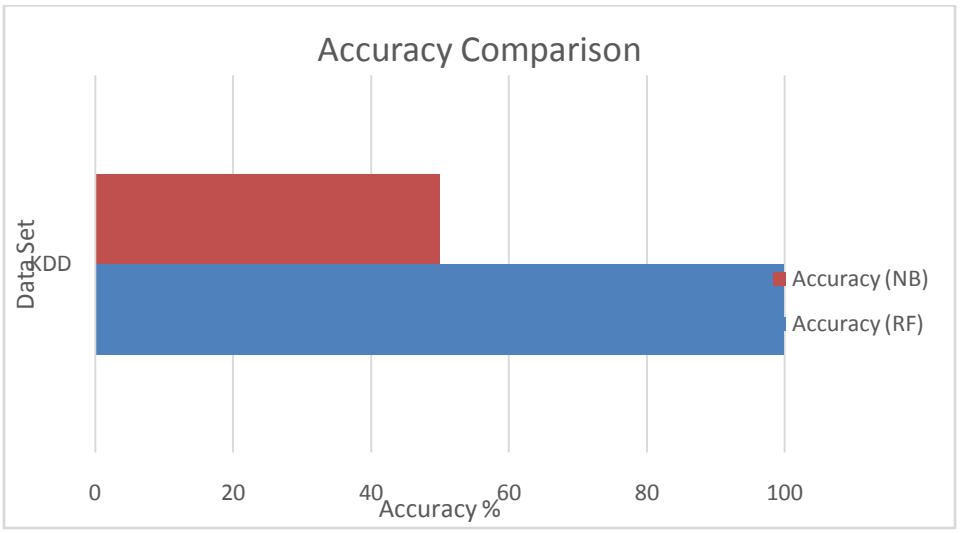

Fig. 2 Accuracy Comparison

V. CONCLUSION

Online intrusion detection is one of the major necessities due to the raise of ecommerce transactions. An effective intrusion detection technique is proposed in this paper which also handles Big Data. The proposed approach uses Random Forest algorithm, an ensemble of several decision trees to build the classifier rules. The usage of several rule creating algorithms provided an added advantage to the algorithm by making it immune to imbalance and missing data. Future research directions of the current system are based on proposing hybrid models to perform effective detection of intrusions in real time.

REFERENCES

[1] J. Anne. "Optimisation, security, privacy and trust in e-business systems." Journal of Computer and System Sciences 81.6: 941-942, 2015.
[2] Z. Richard, T.M. Khoshgoftaar, and R.Wald, 2015. "Intrusion detection and Big Heterogeneous Data: a Survey." Journal of Big Data 2.1: 1-41, 2015.
[3] R. Singh, H. Kumar, and R.K. Singla, "An intrusion detection system using network traffic profiling and online sequential extreme learning machine, Expert Systems with Applications," Volume 42, Issue 22, Pages 8609-8624, 2015.
[4] A. Mauro, M.Colajanni, and M. Marchetti, "A collaborative framework for intrusion detection in mobile networks." Information Sciences,2015.
[5] A.mamory, O.Safaa, and F.S.Jassim, "On the designing of two grains levels network intrusion detection system." Karbala International Journal of Modern Science 1.1: 15-25, 2015.
[6] B. Waleed, A.James, and M.Pannu, 2015. "Improving network intrusion detection system performance through quality of service configuration and parallel technology." Journal of Computer and System Sciences 81.6: 981-999, 2015.
[7] Z. Kai, Z. Cai, X. Zhang, Z. Wang, and B. Yang, "Algorithms to speedup pattern matching for network intrusion detection systems." Computer Communications 62: 47-58,2015.
[8] R. Samaneh, P. Hingston, and C. Lam, "Evolving statistical rulesets for network intrusion detection." Applied Soft Computing 33: 348-359, 2015.
[9] B.H. Monowar,D.K. Bhattacharyya, J.K. Kalita, "An empirical evaluation of information metrics for low-rate and high-rate DDoS attack detection." Pattern Recognition Letters 51: 1-7, 2015.
[10] K. Abdullah and A. Sami, "SysDetect: A systematic approach to critical state determination for Industrial Intrusion Detection Systems using Apriori algorithm." Journal of Process Control,2015.
[11] W. Hui, G. Zhang, E. Mingjie, and N. Sun, "A novel intrusion detection method based on improved SVM by combining PCA and PSO." Wuhan University Journal of Natural Sciences 16, no. 5: 409-413, 2011.
[12] Shittu, Riyanat, Healing, A., Ghanea-Hercock, R., Bloomfield, R. and Rajarajan, M. 2015. Intrusion alert prioritisation and attack detection using post-correlation analysis. Computers & Security 50: 1-15
[13] L.Wei-Chao, S. Ke, and C. Tsai, 2015. CANN: An intrusion detection system based on combining cluster centers and nearest neighbors. Knowledge-Based Systems 78 : 13-21, 2015.
[14] C. Jaeik, T. Shon, K.Choi, and J. Moon, "Dynamic learning model update of hybrid-classifiers for intrusion detection." The Journal of Supercomputing 64, no. 2 : 522-526, 2013.
[15] P.S. Nilkanth, and R.S. Bichkar, "Genetic algorithm with variable length chromosomes for network intrusion detection." International Journal of Automation and Computing: 1-6,2015.