Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Data Analytics Approach to Cyber Crime Underground Economy Using Machine Learning

[1]M.G.Mythili, [2]N.Syed Harif Rezwan, [3]H.Suraj Chandra Gowtham,
[4]M.Satish Kumar Reddy[5], S.Rajashekar Reddy

[1]Assistant professor, School of Computing, Department of Computer Science and Engineering, Bharath Institute of Higher Education And Research, Chennai, India- 600073 .

[2, 3,4,5]Student , , School of Computing, Department of Computer Science and Engineering, Bharath Institute of Higher Education And Research, Chennai, India- 600073.

[1] Sivaraman2006@gmail.com[2]
rezwanharif@gmail.com[3] surajcg989@gmail.com[4] satishkumarreddy2002@gmail.com[5] rajashekarreddy037@gmail.com

Abstract:

A framework for Android malware detection based on permissions is presented. This framework uses multiple linear regression methods. Application permissions, which are one of the most critical building blocks in the security of the Android operating system, are extracted through static analysis, and security analyzes of applications are carried out with machine learning techniques. Based on the multiple linear regression techniques, two classifiers are proposed for permission-based Android malware detection. As a result, remarkable performances are obtained with classification algorithms models without the need for very complex classification algorithms. In our project we have used algorithms like Naïve Bayes (NB) as existing system and Decision Tree (DT) as proposed system is used. All are measured in terms of accuracy. From the results its proved that proposed Decision Tree (DT) works better than existing Naïve Bayes (NB).

Index Terms— Accuracy, Precision, Recall, NB, DT, Malware

I. Introduction

When the first mobile phones were considered, generally speaking or short message transactions were carried out with mobile phones in daily life. However, with mobile phones used today, remarkable transactions such as banking transactions, social media use, and personal data storage take place. Because of these essential processes, mobile devices are the main target of malware developers. Android is an open-source Linux-based mobile operating system. Since it is open-source and free, mobile device manufacturers prefer this operating system on their devices. Therefore, the majority of the market consists of Android devices. According to Statista's data, 30% of the market in the fourth quarter of 2010 consisted of the Android operating system. In the second quarter of 2018, 88% of the market was Android operating systems. In addition to Android being an open-source operating system, it is very flexible for users that applications are provided to devices such as other stores or third-party applications apart from the official application stores. For this reason, Android is frequently preferred by many people around the world.

Although applications from unofficial application repositories or third-party application developers are very advantageous for users, it should not be ignored that some of these applications are malware. Apps in official app repositories are carefully analyzed and published in app repositories. However, malware is common even in official application repositories. In the research conducted by Wang et al., more than 6 million

applications downloaded from 17 application stores are evaluated [3]. While 16 of these stores are widely used in China, the first place is Google Play. In general, it is revealed that Google Play is more reliable than other application stores. However, it is possible to see malware in almost all stores.

II. LITERATURESURVEY

From the last few years, security in wireless sensor network (WSN) is essential because WSN application uses important information sharing between the nodes. There are large number of issues raised related to security due to open deployment of network. The attackers disturb the security system by attacking the different protocol layers in WSN. The standard AODV routing protocol faces security issues when the route discovery process takes place. The data should be transmitted in a secure path to the destination. Therefore, to support the process we have proposed a trust based Malware detection system (NL-MDS) for network layer in WSN to detect the Black hole attackers in the network. The sensor node trust is calculated as per the deviation of key factor at the network layer based on the Black hole attack. We use the watchdog technique where a sensor node continuously monitors the neighbor node by calculating a periodic trust value. Finally, the overall trust value of the sensor node is evaluated by the gathered values of trust metrics of the network layer (past and previous trust values). This NL-MDS scheme is efficient to identify the malicious node with respect to Black hole attack at the network layer. To analyze the performance of NL-MDS, we have simulated the model in MATLAB R2015a, and the result shows that NL-MDS is better than Wang et al. [11] as compare of detection accuracy and false alarm rate.

Safety critical, Internet of Things (IoT) and space-based applications have recently begun to adopt wireless networks based on commercial off the shelf (COTS) devices and standardized protocols, which inherently establishes the security challenge of malicious Malwares. Malicious Malwares can cause severe consequences if undetected, including, complete denial of services. Particularly, any safety critical application requires all services to operate correctly, as any loss can be detrimental to safety and/or privacy. Therefore, in order for these safety critical services to remain operational and available, any and all Malwares need to be detected and mitigated. Whilst Malware detection is not a new research area, new vulnerabilities in wireless networks, especially wireless sensor networks (WSNs), can be identified. In this paper, a specific vulnerability of WSNs is explored, termed here the matched protocol attack. This malicious attack uses protocol-specific structures to compromise a network using that protocol. Through attack exploration, this paper provides evidence that traditional spectral techniques are not sufficient to detect an Malware using this style of attack. Furthermore, a ZigBee cluster head network, which co-exists with ISM band services, consisting of XBee COTS devices is utilized, along with a real time spectrum analyzer, to experimentally evaluate the effect of matched protocol interference on a realistic network model. Results of this evaluation are provided in terms of device errors and spectrum use. This malicious challenge is also examined through Monte-Carlo simulations. A potential detection technique, based on coarse inter-node distance measurements, which can theoretically be used to detect matched protocol interference and localize the origin of the source, is also suggested as a future progression of this work. Insights into how this attack style preys on some of the main security risks of any WSN (interoperability, device

In wireless sensor networks (WSNs), data can be subject to malicious attacks and failures, leading to unreliability. This vulnerability poses a challenge to environmental monitoring applications by creating false alarms. To guarantee a trustworthy system, we therefore need to detect abnormal nodes. In this paper, we propose a new framework for detecting abnormal nodes in clustered heterogeneous WSNs. It makes use of observed spatiotemporal (ST) and multivariate-attribute (MVA) sensor correlations, while considering the background knowledge of the monitored environment. Based on the ST correlations, the collected data is analyzed by computing the crosscorrelation between sensor streams. A new method is proposed for evaluating the intensity of the correlation between two sensor streams. The crosscorrelation value obtained is compared against two thresholds, the lag threshold and the correlation threshold. Based on available background knowledge and the observed MVA correlations, a number of rules are presented to detect abnormal nodes while identifying real events. Our experiments on real-world sensor data demonstrate that our approach captures the correlation and discovers abnormal nodes efficiently.

Wireless sensor networks, due to their nature, are more prone to security threats than other networks. Developments in WSNs have led to the introduction of many protocols specially developed for security purposes. Most of these protocols are not efficient in terms of putting an excessive computational and energy consumption burden on small nodes in WSNs. This paper proposes a knowledge-based context-aware approach for handling the Malwares generated by malicious nodes. The system operates on a knowledge base, located at the base station, which is used to store the events generated by the nodes inside the network. The events are categorized and the cluster heads (CHs) are acknowledged to block maliciously repeated activities generated. The CHs can also get informational records about the maliciousness of intruder nodes by using their inference engines. The mechanism of events logging and analysis by the base station greatly affects the performance of nodes in the network by reducing the extra security-related load on them.

As the medical body sensor network (BSN) is usually resource limited and vulnerable to environmental effects and malicious attacks, faulty sensor data arise inevitably which may result in false alarms, faulty medical diagnosis, and even serious misjudgment. Thus, faulty sensory data should be detected and removed as much as possible before being utilized for medical diagnosis-making. Most available works directly employed fault detection schemes developed in traditional wireless sensor network (WSN) for body sensor fault detection. However, BSNs adopt a very limited number of sensors for vital information collection, lacking the information redundancy provided by densely deployed sensor nodes in traditional WSNs. In light of this, a Bayesian network model-based sensor fault detection scheme is proposed in this paper, which relies on historical training data for establishing the conditional probability distribution of body sensor readings, rather than the redundant information collected from a large number of sensors. Furthermore, the Bayesian network-based scheme enables us to minimize the inaccuracy rate by optimally tuning the threshold for fault detection. Extensive online dataset has been adopted to evaluate the performance of our fault detection scheme, which shows that our scheme possesses a good fault detection accuracy and a low false alarm rate.

Due to the lack of centralized coordination, physical protection, and security requirements of inherent network protocols, wireless sensor networks (WSNs) are vulnerable to diverse denial-of-service (DoS) attacks that primarily target service availability by disrupting network routing protocols or interfering with on-going communications. In this paper, we propose a light-weight countermeasure to a selective forwarding attack, called SCAD, where a randomly selected single checkpoint node is deployed to detect the forwarding misbehavior of malicious node. The proposed countermeasure is integrated with timeout and hop-by-hop retransmission techniques to quickly recover unexpected packet losses due to the forwarding misbehavior or bad channel quality. We also present a simple analytical model and its numerical result in terms of false detection rate. We conduct extensive simulation experiments for performance evaluation and comparison with the existing CHEMAS and CAD schemes. The simulation results show that the proposed countermeasure can improve the detection rate and packet delivery ratio (PDR) as well as reduce the energy consumption, false detection rate, and successful drop rate.

Wireless sensor networks (WSNs) propose the promise of a flexible, low cost solution for monitoring critical infrastructure. Sensor networks have been recommended for applications such as traffic monitoring, military and battlefield surveillance. Wireless sensor networks are more prone to security attacks due to their broadcasting nature of the transmission medium and unattended deployment of nodes in hostile and unfriendly areas where they are not protected as compared to wired networks. Attackers can deploy various types of security attacks to obstruct the security of WSNs. Network layer attacks are more severe since if the routing information is disregarded, disturbances may bring about routing loops, changing of routes etc. Selective forwarding attack is a type of active attack affecting network layers that selectively drops or refuses to forward the data packets. This paper discusses about an energy efficient detection-removal algorithm for effective detection of selective forwarding attack in a clustered WSN scenario. The impact of the malicious node in network parameters like packet delivery ratio, throughput, residual energy of network and end to end delay are analyzed.

Wireless Sensor Networks are extensively used in developing applications for surveillance, habitat monitoring, border security, Malware detection etc. Most of these applications require secure data transmission among the nodes of the network. Out of the different types of attacks a data critical application faces, False Data Injection attacks are the most damaging one. So prevention of False Data Injection attacks is a crucial aspect while building data critical wireless sensor network applications. Researchers have suggested cryptographic schemes like RSA, ECC for the prevention of False Data Injection Attacks. Use of cryptographic techniques increases the computation complexity on all the nodes and the energy constraints on WSN demands an alternate solution for False Data Injection attacks prevention. The proposed work aims on using trust parameter of every nodes to distinguish malicious and non-malicious nodes and use only trusted nodes to forward the packet to destination thus by prevention FDI attacks. Simulation is carried out with the help of Network Simulator 2 (NS2). The results shows the energy consumption is less in the proposed scheme compared to the cryptographic technique

III EXISTINGSYSTEM

Naïve Bayes (NB):

In statistics, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels.

Naïve Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

In the statistics and computer science literature, naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes. All these names reference the use of Bayes' theorem in the classifier's decision rule, but naïve Bayes is not (necessarily) a Bayesian method.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable. For example, a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

For some types of probability models, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

IV. METHODOLOGY

The IDS have been implemented in organizations to collect and analyze various types of attacks within a host system or a network. In addition, to identify and detect possible threats violations, which involve both intrusions, which are the attacks from outside the organizations and misuses that are known as the attacks within the organizations. In this paper, we proposed the integrated model which involves a combination of the two systems Intrusion Detection (ID) and Intrusion Prevention (IP) adding to those getting benefits from well-known techniques: intruder Detection (ID) which is totally different from most of the recent works that focused only on using one system, either detection or prevention and also using either Intruder detection or Signature based detection. Some works even used a hybrid method which is a combination of both such as the work presented where the researchers used ID based on Signature but even then, their method was not provided with prevention capabilities.

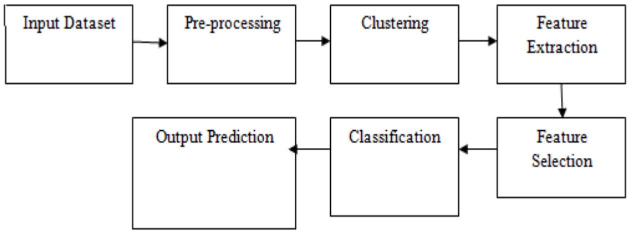

Fig 1: Block Diagram

MODULE DESCRIPTION

1. Input dataset
2. Analysis of size of data set.
3. Oversampling.
4. Training and Testing.
5. Apply algorithms.
6. Predict results.

1. Input dataset:
Dataset can be taken from online data source provider from the internet sources. We have to collect a huge dataset in volume so as to predict the accuracy in an efficient manner.

2. Analysis of data set:
Here the analysis if dataset takes place. The size of data is taken into consideration for the data process.

3. Oversampling (Using SMOTE): we have created a detailed history of all malware attack as dataset that is been happened over a long duration.

4. Training and Testing Subset: As the dataset is imbalanced, many classifiers show bias for majority classes. The features of minority class are treated as noise and are ignored. Hence it is proposed to select a sample dataset.

5. Applying algorithm: Following are the classification algorithms used to test the sub-sample dataset.
a. Decision Tree (DT) and
b. Naïve Bayes (NB)

6. Predicting results: The test subset is applied on the trained model. The metrics used is accuracy. The accuracy Curve is plotted and the desirable results are achieved.

V. TOOLSUSED

OpenCV is a storage of programming operations for actual time computer scope actually created by Intel and now assisted by Willogarage. It is liberately used under the free source BSD license. This has greater than five hundred effective set of rules to be followed. It is widely employed around the world, with forty thousand users in the user community. Used in wide limit ranging from communicating resource, to fine audit, and upcoming robotics. The package is developed in C, which does it movable to few particular surface such as Digital Signal Processor. Packaging for languages such as C, Python, Ruby and Java (using JavaCV)

A. Python

Python is a remarkably powerful dynamic, object-oriented programming language that is used in a wide variety of application domains. It offers strong support for integration with other languages and tools, and comes with extensive standard libraries. To be precise, the following are some distinguishing features of Python:

- Very clear, readable syntax.
- Strong introspection capabilities.
- Full modularity.
- Exception-based error handling.
- High level dynamic data types.
- Supports object oriented, imperative and functional programming styles.
- Embeddable.
- Scalable
- Mature

With so much of freedom, Python helps the user to think problem centric rather than language centric as in other cases. These features makes Python a best option for scientific computing.

B. Open CV

Open CV is a library of programming functions for real time computer vision originally developed by Intel and now supported by Willogarage. It is free for use under the open source BSD license. The library has more than five hundred optimized algorithms. It is used around the world, with forty thousand people in the user group. Uses range from interactive art, to mine inspection, and advanced robotics. The library is mainly written in C, which makes it portable to some specific platforms such as Digital Signal Processor. Wrappers for languages such as C, Python, Ruby and Java (using Java CV) have been developed to encourage adoption by a wider audience. The recent releases have interfaces for C++. It focuses mainly on real-time image processing. Open CV is a cross-platform library, which can run on Linux, Mac OS and Windows. To date, Open CV is the best open source computer vision library that developers and researchers can think of.

C. Tesseract

Tesseract is a free software OCR engine that was developed at HP between 1984 and 1994. HP released it to the community in 2005. Tesseract was introduced at the 1995 UNLV Annual Test OCR Accuracy and is currently developed by Google released under the Apache License. It can now recognize 6 languages, and is fully UTF8 capable. Developers can train Tesseract with their own fonts and character mapping to obtain perfect efficiency.

VI SIMULATION RESULT

The Accuracy graph is plotted between Naïve Bayes and Decision Tree. Accuracy graph is given in the following figure 2.

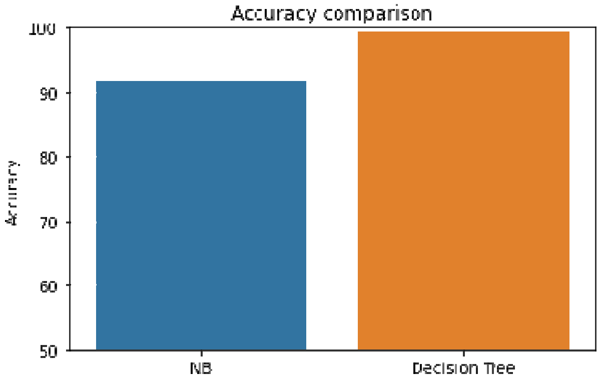

Fig 2: Accuracy Analysis

VII CONCLUSION

Application permissions are significant in Android operating system security. These permissions, which are extracted from applications, are used as attributes to detect malicious software with machine learning algorithms in this study. Android malware detection is carried out with two rule-based classification models using Hybrid models. However, classifiers are quite simple and easy to use. This is the most significant advantage of the proposed approaches. Future research should consider other machine learning algorithms to ascertain more efficient ways to perform the classification technique on the datasets. It is recommended that further research should be carry out on other parameters that can improve the accuracy of detection.

REFERENCES

[1] I. F. Akyildiz et al., “Wireless Sensor Networks: A Survey, “Elsevier Comp. Networks, vol. 3, no. 2, 2019, pp. 393–422

[2] G.Li, J.He, Y. Fu. “Group-based Malware detection system in wireless sensor networks” Computer Communications, Volume 31, Issue 18 (December 2019)

[3] Michael Brownfield, “Wireless Sensor Network Denial of Sleep Attack”, Proceedings of the 2019 IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY.

[4] FarooqAnjum, DhanantSubhadrabandhu, SaswatiSarkar *, Rahul Shetty, “On Optimal Placement of Malware Detection Modules in Sensor Networks”, Proceedings of the First International Conference on Broadband Networks (BROADNETS19).

[5] Parveen Sadotra et al, International Journal of Computer Science and Mobile Computing, Vol.5 Issue.9, September-2019, pg. 23-28

[6] K. Akkayaand M. Younis, ―A Survey of Routing Protocols in Wireless Sensor Networks, ‖ in the Elsevier Ad Hoc Network Journal, Vol. 3/3 pp. 325-349, 2019.

[7] A. Abduvaliyev, S. Lee, Y.K Lee, “Energy Efficient Hybrid Malware Detection System for Wireless Sensor Networks”, IEEE International Conference on Electronics and Information Engineering, Vol.2, pp. 25-29, August 2019.

[8] Parveen Sadotra and Chandrakant Sharma. A Survey: Intelligent Malware Detection System in Computer Security. International Journal of Computer Applications 151(3):18-22, October 2019.

[9] A. Araujo, J. Blesa, E. Romero, D. Villanueva, “Security in cognitive wireless sensor networks. Challenges and open problems”, EURASIP Journal on Wireless Communications and Networking, February 2019.

[10] A. Becher, Z. Benenson, and M. Dorsey, \Tampering with motes: Real-world physical attacks on wireless sensor networks." in SPC (J. A. Clark, R. F. Paige, F. Polack, and P. J.Brooke, eds.), vol. 3934 of Lecture Notes in Computer Science, pp. 104{118, Springer, 2019.

[11] I. Krontiris and T. Dimitriou, \A practical authentication scheme for in-network programming in wireless sensor networks," in ACM Workshop on Real- World Wireless Sensor Networks, 2019.

[12] M. Ali Aydın *, A. HalimZaim, K. GokhanCeylan “A hybrid Malware detection system design for computer network security” Computers and Electrical Engineering 35 (2019) 517–526.