ROLE OF DATA MINING IN MALWARE DETECTION

SHAIK MOHAMMED SHAFIULLA
Assistant Professor,
Department of CSE,
SCIENT INSTITUTE OF TECHNOLOGY,
Hyderabad, Telangana,[INDIA].

VANAM GOPINATH
Assistant Professor,
Department of CSE,
SCIENT INSTITUTE OF TECHNOLOGY,
Hyderabad, Telangana,[INDIA].

ABSTRACT

Data mining is the process of identifying patterns in large datasets. Data mining techniques are heavily used in scientific research (in order to process large amounts of raw scientific data) as well as in business, mostly to gather statistics and valuable information to enhance customer relations and marketing strategies. Data mining has also proven a useful tool in cyber security solutions for discovering vulnerabilities and gathering indicators for baselining.In this paper we discussed about the role of role of data mining in information security and the malware detection process and overview of the Minnesota Intrusion Detection System (MINDS), which uses a suite of data mining based algorithms to address different aspects of cyber security

KEYWORDS : Data mining,malware detection,summarization

I INTRODUCTION

Data mining is the exploration and analysis of large data to discover meaningful patterns and rules. It's considered a discipline under the data science field of study and differs from predictive analytics because it describes historical data, while data mining aims to predict future outcomes. Additionally, data mining techniques are used to build machine learning (ML) models that power modern artificial intelligence (AI) applications such as search engine algorithms and recommendation systems.It is a process that involves analyzing information, predicting future trends, and making proactive, knowledge-based decisions based on large datasets.

While the term data mining is usually treated as a synonym for Knowledge Discovery in Databases (KDD), it's actually just one of the steps in this process. The main goal of KDD is to obtain useful and often previously unknown information from large sets of data.The entire KDD process includes four steps:

Pre-processing – selecting, cleaning, and integrating data

Transformation – transforming information and consolidating it into forms appropriate for mining.Mining – collecting, extracting, analyzing, and statistically processing data

Pattern evaluation – identifying new and unusual patterns and presenting the knowledge gained from data mining

Data mining helps you find new interesting patterns, extract hidden (yet useful and valuable) information, and identify unusual records and dependencies from large databases. To obtain valuable knowledge, data mining uses methods from

statistics, machine learning, artificial intelligence (AI), and database systems.In recent years, many IT industry giants such as Comodo, Symantec, and Microsoft have started using data mining techniques for malware detection.

II DATA MINING METHODS

Many methods are used for mining big data, but the following eight are the most common:

Association rules help find possible relations between variables in databases, discover hidden patterns, and identify variables and the frequencies of their occurrence. Classification breaks a large dataset into predefined classes or groups. Clustering helps identify data items that have similar characteristics and understand similarities and differences among data. The decision tree technique creates classification and regression models in the form of a tree structure. The neural network technique is used to model complex relationships between inputs and outputs and to discover new patterns. Regression analysis is used for predicting the value of one item based on the known value of other items in a dataset by building a model of the relationship between dependent and independent variables.

Statistical techniques help find patterns and build predictive models.

Visualization discovers new patterns and shows the results in a way that is comprehensible for users.

III DATA MINING FOR MALWARE DETECTION

Data mining is one of the four detection methods used today for detecting malware. The other three are scanning, activity monitoring, and integrity checking. When building a security app, developers use data mining methods to improve the speed and quality of malware detection as well as to increase the number of detected zero-day attacks.

3.1 Malware detection strategies

There are three strategies for detecting malware:

- Anomaly detection
- Misuse detection
- Hybrid detection

Anomaly detection involves modeling the normal behavior of a system or network in order to identify deviations from normal usage patterns. Anomaly-based techniques can detect even previously unknown attacks and can be used for defining signatures for misuse detectors.The main problem with anomaly detection is that any deviation from the norm, even if it is a legitimate behavior, will be reported as an anomaly, thus producing a high rate of false positives.Misuse detection, also known as signature-based detection, identifies only known attacks based on examples of their signatures. This technique has a lower rate of false positives but can't detect zero-day attacks.A hybrid approach combines anomaly and misuse detection techniques in order to increase the number of detected intrusions while decreasing the number of false positives. It doesn't build any models, but instead uses information from both harmful and clean programs to create a classifier – a set of rules or a detection model generated by the data mining algorithm. Then the anomaly detection system searches for deviations from the normal profile and the misuse detection

system looks for malware signatures in the code.

3.2 Detection process

When using data mining, malware detection consists of two steps:

- Extracting features
- Classifying/clustering

In the first step, various features such as API calls, n-grams, binary strings, and program behaviors are extracted statically and dynamically to capture the characteristics of the file samples. Feature extraction can be performed by running static or dynamic analysis (with or without actually running potentially harmful software). A hybrid approach that combines static and dynamic analysis may also be used.

During classification and clustering, file samples are classified into groups based on feature analysis. To classify samples, you can use classification or clustering techniques.To classify file samples, you need to build a classification model (a classifier) using classification algorithms such as RIPPER, Decision Tree (DT), Artificial Neural Network (ANN), Naive Bayes (NB), or Support Vector Machines (SVM). Clustering is used for grouping malware samples that have similar characteristics.Using machine learning techniques, each classification algorithm constructs a model that represents both benign and malicious classes. Training a classifier using such file sample collection makes it possible to detect even newly released malware.Effectiveness of data mining techniques for malware detection critically depends on the features you extract and the categorization techniques you use.

IV DATA MINING FOR INTRUSION DETECTION

Aside from detecting malware code, data mining can be effectively used to detect intrusions and analyze audit results to detect anomalous patterns. Malicious intrusions may include intrusions into networks, databases, servers, web clients, and operating systems.There are two types of intrusion attacks you can detect using data mining methods.Host-based attacks, when the intruder focuses on a particular machine or a group of machines.Network-based attacks, when the intruder attacks the entire network (for instance, causing a buffer overflow

To detect host-based attacks, you need to analyze features extracted from programs, while to detect network-based attacks, you need to analyze network traffic. And just like with malware detection, you can look for either anomalous behavior or cases of misuse.

V DATA MINING FOR FRAUD DETECTION

You can detect various types of fraud using data mining techniques, from financial fraud to telecommunications fraud and computer intrusions. Fraudulent activities can be detected with the help of supervised and unsupervised learning.With supervised learning, all available records are classified as either fraudulent or non-fraudulent. This classification is then used for training a model to detect possible fraud. The main drawback of this method is its inability to detect new types of attacks. Unsupervised learning methods help identify privacy and security issues in data without using statistical analysis.

VI PROPOSED APPROACH

The MINDS(Minnesota Intrusion Detection System) suite contains various modules for collecting and analyzing massive amounts of network traffic. Typical analyses include behavioral anomaly detection, summarization, scan detection and profiling. Additionally, the system has modules for feature extraction and filtering out attacks for which good signatures have been learnt [8]. Each of these modules will be individually described in the subsequent sections. Independently, each of these modules provides key insights into the network. When combined, which MINDS does automatically, these modules have a multiplicative affect on analysis. As shown in the figure, MINDS system is involves a network analyst who provides feedback to each of the modules based on their performance to fine tune them for more accurate analysis While the anomaly detection and scan detection modules aim at detecting actual attacks and other abnormal activities in the network traffic, the profiling module detects the dominant modes of traffic to provide an effective profile of the network to the analyst. The summarization module aims at providing a concise representation of the network traffic and is typically applied to the output of the anomaly detection module to allow the analyst to investigate the anomalous traffic in very few screenshots.

The various modules operate on the network data in the NetFlow format by converting the raw network traffic using the flow-tools library 2 . Data in NetFlow format is a collection of records, where each record corresponds to a unidirectional flow of packets within a session. Thus each session (also referred to as a connection) between two hosts comprises of two flows in opposite directions. These records are highly compact containing summary information extracted primarily from the packet headers. This information includes source IP, source port, destination IP, destination port, number of packets, number of bytes and timestamp. Various modules extract more features from these basic features and apply data mining algorithms on the data set defined over the set of basic as well as derived features.MINDS is deployed at the University of Minnesota, where several hundred million network flows are recorded from a network of more than 40,000 computers every day.

MINDS is also part of the Interrogator [15] architecture at the US Army Research Labs Center for Intrusion Monitoring and Protection (ARL-CIMP), where analysts collect and analyze network traffic from dozens of Department of Defense sites [7]. MINDS is enjoying great operational success at both sites, routinely detecting brand new attacks that signature-based systems could not have found. Additionally, it often discovers rogue communication channels and the exfiltration of data that other widely used tools such as SNORT [19] have had difficulty identifying.

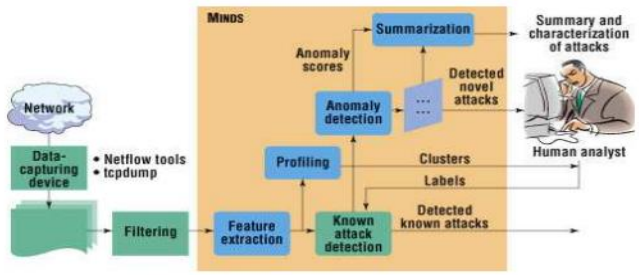

Fig. 1. The Minnesota Intrusion Detection System (MINDS)

VII CONCLUSION

Data mining has great potential as a malware detection tool. It allows you to analyze huge sets of information and extract new knowledge from it.The main benefit of using data mining techniques for detecting malicious software is the ability to identify both known and zero-day attacks. However, since a previously unknown but legitimate activity may also be marked as potentially fraudulent, there's the possibility for a high rate of false positives. MINDS is a suite of data mining algorithms which can be used as a tool by network analysts to defend the network against attacks and emerging cyber threats.

REFERENCES

[1]Rakesh Agrawal, Tomasz Imieliski, and Arun Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data, pages 207–216. ACM Press, 1993.

[2]Daniel Barbara and Sushil Jajodia, editors. Applications of Data Mining in Computer Security. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[3]Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and J Sander. Lof: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pages 93–104. ACM Press, 2000.

[4]Varun Chandola and Vipin Kumar. Summarization – compressing data into an informative representation. In Fifth IEEE International Conference on Data Mining, pages 98–105, Houston, TX, November 2005.

[5]William W. Cohen. Fast effective rule induction. In International Conference on Machine Learning (ICML), 1995.

[6]Dorothy E. Denning. An intrusion-detection model. IEEE Trans. Softw. Eng., 13(2):222–232, 1987.

[7]Eric Eilertson, Levent Ert¨oz, Vipin Kumar, and Kerry Long. Minds – a new approach to the information security process. In 24th Army Science Conference. US Army, 2004.

[8]Levent Ert¨oz, Eric Eilertson, Aleksander Lazarevic, Pang-Ning Tan, Vipin Kumar, Jaideep Srivastava, and Paul Dokas. MINDS - Minnesota Intrusion Detection System. In Data Mining - Next Generation Challenges and Future Directions. MIT Press, 2004.

[9]Levent Ertoz, Michael Steinbach, and Vipin Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In Proceedings of 3rd SIAM International Conference on Data Mining, May 2003.

[10]Anil K. Jain and Richard C. Dubes. Algorithms for Clustering Data. PrenticeHall, Inc., 1988.

[11]Jaeyeon Jung, Vern Paxson, Arthur W. Berger, and Hari Balakrishnan. Fast portscan detection using sequential hypothesis testing. In IEEE Symposium on Security and Privacy, 2004.

[12]Vipin Kumar, Jaideep Srivastava, and Aleksander Lazarevic, editors. Managing Cyber Threats–Issues, Approaches and Challenges. Springer Verlag, May 2005.

[13]Aleksandar Lazarevic, Levent Ert¨oz, Vipin Kumar, Aysel Ozgur, and Jaideep Srivastava. A comparative study of anomaly detection schemes in network intrusion

detection. In SIAM Conference on Data Mining (SDM), 2003.

[14]C. Lickie and R. Kotagiri. A probabilistic approach to detecting network scans.In Eighth IEEE Network Operations and Management, 2002.

[15]Kerry Long. Catching the cyber-spy, arl's interrogator. In 24th Army Science Conference. US Army, 2004.

[16]V. Paxon. Bro: a system for detecting network intruders in real-time. In Eighth IEEE Network Operators and Management Symposium (NOMS), 2002.

[17]Phillip A. Porras and Alfonso Valdes. Live traffic analysis of tcp/ip gateways. In NDSS, 1998.

[18]Seth Robertson, Eric V. Siegel, Matt Miller, and Salvatore J. Stolfo. Surveillance detection in high bandwidth environments. In DARPA DISCEX III Conference, 2003.

[19]Martin Roesch. Snort: Lightweight intrusion detection for networks. In LISA, pages 229–238, 1999.

[20]Gyorgy Simon, Hui Xiong, Eric Eilertson, and Vipin Kumar. Scan detection: A data mining approach. Technical Report AHPCRC 038, University of Minnesota – Twin Cities, 2005.

[21]Gyorgy Simon, Hui Xiong, Eric Eilertson, and Vipin Kumar. Scan detection: A data mining approach. In Proceedings of SIAM Conference on Data Mining (SDM), 2006.

[22]Anoop Singhal and Sushil Jajodia. Data mining for intrusion detection. In Data Mining and Knowledge Discovery Handbook, pages 1225–1237. Springer, 2005.

[23]Stuart Staniford, James A. Hoagland, and Joseph M. McAlerney. Practical automated detection of stealthy portscans. Journal of Computer Security, 10(1/2):105–136, 2002.

[24]Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Introduction to Data Mining. Addison-Wesley, May 2005.

[25]Nicholas Weaver, Stuart Staniford, and Vern Paxson. Very fast containment of scanning worms. In 13th USENIX Security Symposium, 2004.