ISSN 2616-5775 Vol. 7, Issue 1: 72-77, DOI: 10.25236/AJCIS.2024.070111

Research on Malicious Code Detection Techniques Based on Data Mining and Machine Learning

Ye Zhang

University of Pittsburgh, Pittsburgh, United State
yez12@pitt.edu

Abstract: *With the increasing number and types of malicious codes and the development of code confusion technology, malicious code detection becomes more and more difficult. The traditional malicious code monitoring technology has been unable to realize the accurate identification of malicious code, which leads to the increasing risk of the computer system being threatened by malicious code. Based on the current development trend of malicious code, this paper proposes a malicious code monitoring technology through data mining and machine learning, which can realize a more accurate identification of malicious code and create better conditions for the detection and killing of malicious code.*

Keywords: *malicious code; detection technology; data mining; machine learning*

1. Introduction

Malicious code is a kind of program code, the impact on the computer mainly includes causing computer malfunction, information leakage, destroying computer data as well as affecting the normal operation of the computer system use. It has become one of the important forms of threatening computer security. Relevant research shows that the network security incidents caused by malicious code are increasing year by year, with an annual increase of more than 50%. Malicious code has caused huge economic losses. Due to the relatively large number of malicious code, and new malicious code is constantly appearing, resulting in the detection speed and efficiency of traditional malicious code monitoring technology can not meet the current demand for malicious code, there is an urgent need for a more efficient malicious code monitoring technology to protect the security of the computer system.

2. Data mining and machine learning overview

2.1 Overview of data mining

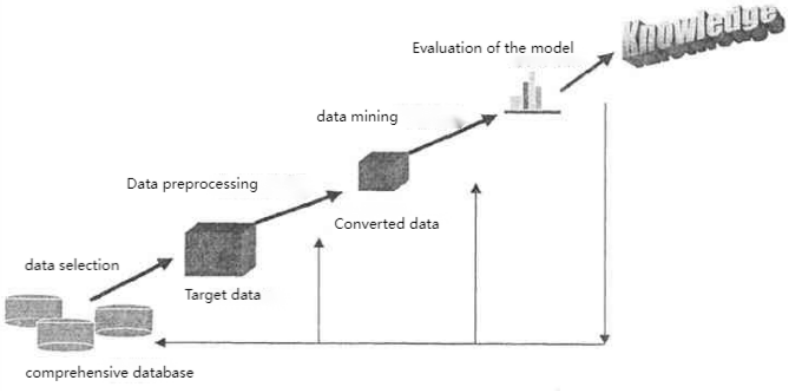

Fig. 1 Process diagram of data mining

Data mining refers to the complex process of extracting and mining knowledge such as unknown and valuable patterns or laws from a large amount of data. The process of data mining is shown in Figure 1. (1) Data Cleaning. Data cleaning is used to remove the noise in the data and data completely unrelated to the topic; (2) Data integration. Aggregate relevant data from multiple data sources through

ISSN 2616-5775 Vol. 7, Issue 1: 72-77, DOI: 10.25236/AJCIS.2024.070111

data integration; (3) Data transformation. Transform the data into a data storage form that is easy to be processed by data mining; (5) Pattern evaluation, screen the mined data according to certain evaluation criteria, and select the meaningful pattern knowledge; (6) Knowledge representation, display the mined knowledge to the user through visualization technology and knowledge expression technology.

Data mining technology is the effective integration of database, information retrieval, statistics, machine learning, algorithms and other fields of theory and technology to achieve the relevant behaviour of judgment and prediction, so as to provide more effective support for people's decision-making. The current typical data mining methods mainly include classification analysis, cluster analysis, association rule analysis and sequence pattern analysis. First of all, classification analysis refers to the mapping of data into pre-designed groups or categories, and then assign the data to different categories according to the attributes of the data through classification functions or classification models. In other words, the type of data can be determined by analysing the attributes of the data and finding out the model of the data attributes. In this way, it is possible to analyse the existing data information to determine which type the new data belongs to. Next is cluster analysis. Cluster analysis also groups the data, but the groups of cluster analysis are not predetermined, but based on the characteristics of the data to find the similarity of the data, and then the definition of the ancestors. In the process of cluster analysis, data will be divided into multiple meaningful sub-sets based on the degree of similarity measured. This division aims to group the data in each sub-set with strong similarity, while maintaining a certain degree of similarity between different sets of data. Therefore, in cluster analysis, how to effectively define the similarity of data is a very key element. The third is association rule analysis. There are certain interrelationships between different data, but these interrelationships will not be directly reflected in the data, through the association rule analysis to determine the interrelationships between the data, you can build a hidden network of associations, so that you can better describe the closeness of the relevant data and relationships. However, in some cases, the precise correlation function between the data is not known, which leads to a certain level of confidence in the correlation rules formed in the process of correlation rule analysis, and the level of confidence will limit the strength of the correlation rules. The fourth is sequence pattern analysis. Sequential pattern analysis and association rule analysis have a certain degree of similarity, both are to analyse the links between the data. But serial pattern analysis focuses more on the time-related connections between data. So when data relationship description is performed through series pattern analysis, the data of each series is a set of data set arranged according to time.

2.2 Overview of machine learning

Machine learning is the acquisition of new knowledge and skills through computer simulation or implementation of human learning behavior. Then the existing knowledge is reorganized to achieve further improvement of its own performance. So machine learning is an important branch of artificial intelligence. The machine learning model is shown in Figure 2. Through the environment to provide the system learning unit with the appropriate external information sources, and then the system unit will be based on this information to modify the knowledge base, so as to increase the efficiency of the entire system to perform the task. The execution unit executes the task using the relevant knowledge stored in the knowledge base and provides feedback to the learning unit for further learning input.

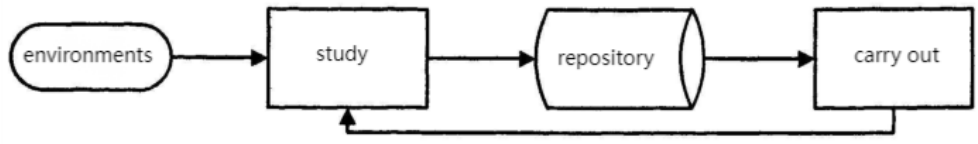

Fig. 2 Machine learning model

The more common learning methods for machine learning include mechanical learning, inductive learning, analogue learning, explanation-based learning, and neural network-based learning. The first is mechanical learning. Mechanical learning is one of the easiest methods of machine learning, by coarsely storing new knowledge and retrieving it when it needs to be used. This process does not require computation or reasoning. Once the mechanical learning system executes a partially solved problem, the whole system keeps a record of the problem and the solution strategy, and then after encountering the problem again, it is able to perform the problem directly. Mechanical learning is a learning method based on memory and retrieval, and the learning method is relatively simple. However, in the process of practical application, the system needs to be able to store information in an organised way, and be able to combine information and control the direction of retrieval. Therefore, the use of

ISSN 2616-5775 Vol. 7, Issue 1: 72-77, DOI: 10.25236/AJCIS.2024.070111

mechanical learning needs to pay attention to the storage and organisation of information, the stability of the environment and the applicability of the stored information and the calculation of the trade-offs between the issues. Next is inductive learning. Inductive learning is a symbolic learning method that represents the process of conceiving hypotheses from examples. In the process of inductive learning, learners reason inductively from facts provided or hypotheses observed to come to a certain concept. So inductive learning is the process of reasoning from part to whole and particular to half. In machine learning, the inductive learning problem is described as a problem of guided search for general rules using training examples. The whole possible instances are constructed into an instance space, the whole possible general rules constitute a rule space, and the task of learning is to search the instance space and the rule space simultaneously and coordinately. The third is analogy learning, which is a reasoning method that provides a clear and concise description of the similarity of objects. It is also capable of transferring the task of testing similarity properties from the speaker to the listener. (1) Representation of analogical learning: assuming that the knowledge of the object is soared to be a set of frames, analogical learning transfers the value of a slot from one frame to the slot of another frame, and generates a recommendation slot through the source frame to achieve the transfer to the target frame, and at the same time, through the use of the existing information in the target frame for similarity screening. (2) Analogical learning solution: after completing the slots of the source frame to establish the possible transmission frames, it is necessary to screen them through the knowledge of the target frame.A,Select which slots are not filled on the target frame; B,Select the slots of the typical instances in the target frame; C,Select the slots which are closely related to the target; D,Select slots which are similar to the target slots; ESelect slots similar to the slots which are closely related to the target of the slot. In this way, the objects to be compared can be represented by the framework, and the close relationship of the objects to be compared can be found according to the ISA hierarchical structure. From there, the generation of possible analogues and the selection of the best are performed based on the generation test method. The fourth is explanation-based learning. Explanation-based VMs bring together a large number of results into a unified and simple framework to analyse why examples are concrete examples of the target concept. At the same time, the components related to the concrete examples are eliminated and a description of the target concept is generated. In this way, abstract goal concepts can be made concrete and simpler to understand and manipulate. This description of the target concept can then provide some experience in solving similar problems. Explanation-based learning can be divided into two stages. The first stage is analysis, which proves that the instances are instances of the target concept through the generation of the proof tree, and the second stage is the explanation-based generalisation stage, which achieves the generalisation of the explanation by substituting the common variables of the proof tree species to form an explanation-based generalisation tree, and finally obtains the sufficient conditions of the target concept. Finally, there is neural network based learning. Neural networks work by learning and using these two non-linear processes, essentially, neural network learning is a form of inductive learning, which stabilises the network in a particular state by running iteratively on a large number of instances and making constant modifications to the distribution of weights through an internal adaptive process. In neural networks, the specific gist and knowledge obtained from learning can be represented by a large number of interconnected structures and connected weight distributions. Thus, in a particular input pattern, the neural network generates an output pattern by forward computation and at the same time obtains the logical concepts represented by the nodes. The output signals are then compared and analysed to obtain a particular solution.

3. Malicious code recognition technology based on data mining and machine learning

3.1 Malicious code detection architecture based on data mining and machine learning

The algorithmic architecture of malicious code detection technology based on data mining and machine learning proposed in this paper is shown in Figure 3. The whole detection algorithm consists of two phases, training and testing, and the training phase is used to complete the training related to static disassembly of samples, feature extraction and selection, and integrated classifier construction. Through static disassembly to determine the shelling situation and type of malicious code, and then through the appropriate shelling program to shell the malicious code. The structural features, byte sequences and instruction sequences of the code are extracted through feature extraction. This will facilitate the learning of subsequent classification algorithms. In this process, feature reduction and redundant feature parsimony are required for high dimensional features. The main role of the integrated classifier is to perform the construction of multiple classification training, selection and integration processes. And the main role of the testing phase is to complete the testing of samples.

ISSN 2616-5775 Vol. 7, Issue 1: 72-77, DOI: 10.25236/AJCIS.2024.070111

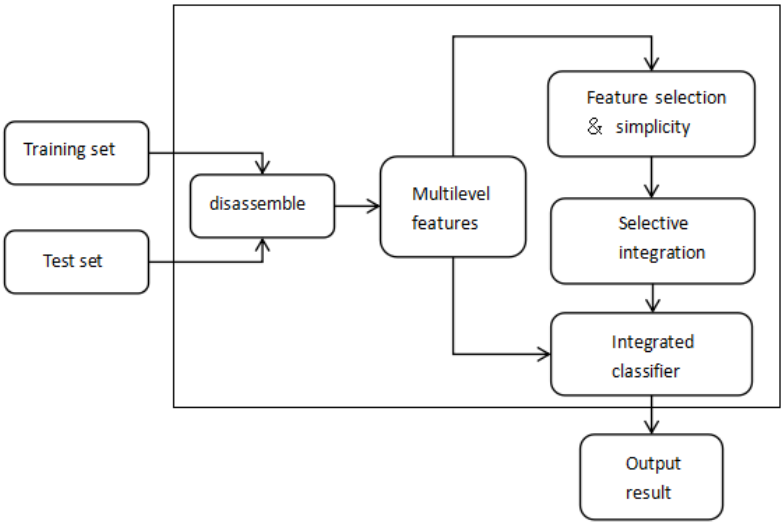

Fig. 3 Algorithmic architecture of malicious code detection technology

3.2 Code Preprocessing

The source data for this study used the sample data from VX Heavens, and a total of 25,584 malicious program samples in PE format were downloaded. Each malicious program sample is composed of DOS file header, DOS block, PE file header, section table, code, data, and resource saving. The malicious code consists of Virus, Trojan, Worm, and Rootkit. In this study, 2768 code samples were randomly selected from each category of malicious code using Rootkit as the benchmark, and the malicious code was classified and recognized by cross-validation.

3.3 Code feature extraction

The first is the text structure feature. Text structure feature refers to the static structure information of PE file code, which mainly includes the normal situation of the entry point, PE header byte entropy, and the number of standard sections. A 19-dimensional code feature vector is obtained by analyzing the PE file structure of the malicious code. Then the features of each dimension of the malicious code are represented by Boolean or numerical values. Next is the byte layer features. The byte-level features of malicious code represent the regular characteristics of the code in the computer binary storage sequence. In this paper, the Hexview tool is used to convert each malicious code into a hexadecimal byte sequence, and then the byte features of the code byte sequence are obtained by using the n-gram sliding window. Setting the sliding length of the n-gram window to 2 in this process can effectively avoid the situation that the feature dimension of the byte layer is too long resulting in the overflow of the system content. The feature dimension of the byte layer is set to 100,000 to perform the feature statistics of the existing byte sequences. The third is the command layer feature. The instruction layer features of malicious code refer to the operation and operand sequence features of the code in the process of executing instructions. The relationship between code, opcode and operand is reflected by the way of defining the code program and operand sequence. In this paper, IDA Pro6.1 disassembly tool is used to realize the reverse compilation of each code sample, and then the n-gram algorithm is used to manipulate the compilation results to obtain the characteristics of the common instruction sequences of the code. In this way 5112 dimensional features were obtained from the code and the sequential frequency of occurrence of these features in the sample instructions was determined.

3.4 Code Feature Selection

In this paper, the main method of principal component analysis is used to realize the dimension reduction of the feature matrix combined with the structure layer, byte layer and code layer of the code. Then the feature matrix obtained after dimensionality reduction is used as the input of the classifier, so as to obtain the classification and recognition results of malicious code through the evaluation of feature combinations. In this way, the covariance matrix of the feature matrices of the text structure

ISSN 2616-5775 Vol. 7, Issue 1: 72-77, DOI: 10.25236/AJCIS.2024.070111

layer, the byte layer and the code layer can be used to obtain the contribution rate of different feature values. Then the selection of code features is done based on the contribution rate of different features.

3.5 Classification and Recognition of Malicious Code

Firstly, the performance metrics for the classification of malicious code need to be determined. This bit uses a variety of classifiers to find the most effective malicious code classification method, and evaluates the malicious code classification method by three evaluation indexes: accuracy, sensitivity and specificity. The accuracy reflects the ability of the classifier to determine the sample set; the sensitivity reflects the ability of the classifier to predict positive samples into positive samples; and the specificity reflects the ability of the classifier to predict negative samples into negative samples. The samples of malicious code are then classified and recognized through the WEKA data mining platform using five classification methods: NaiveBayes, J48, JRip, SVM, and KNN. Thus, the classification results of different classifiers are obtained, as shown in Table 1.

Table 1 Classification results of different features in different classifiers

classification algorithm	Evaluation indicators	structural characteristic	Byte Characteristics	Instruction features	Portfolio Characteristics
NavieBayes	accuracy	0.7361	0.9765	0.8514	0.9386
	sensitivity	0.7543	0.9817	0.8762	0.9435
	specificity	0.7157	0.9223	0.8341	0.8936
J48	accuracy	0.6755	0.9234	0.9347	0.9463
	sensitivity	0.7062	0.9321	0.9469	0.9517
	specificity	0.6564	0.8992	0.8643	0.9072
JRiP	accuracy	0.6713	0.8977	0.9368	0.9365
	sensitivity	0.6817	0.9054	0.9452	0.9386
	specificity	0.6525	0.8776	0.8964	0.9089
SVM	accuracy	0.6675	0.9742	0.98141	0.9255
	sensitivity	0.6843	0.9855	0.9821	0.9364
	specificity	0.6137	0.9243	0.9157	0.8861
KNN	accuracy	0.7375	0.8076	0.9114	0.9412
	sensitivity	0.7384	0.8527	0.9215	0.9537
	specificity	0.6961	0.7932	0.8773	0.9185

Through Table 1, it can be found that in the single code feature classification experiment, the accuracy of malicious code identification through structural features is the lowest, and the accuracy of malicious code identification through instruction features is the highest. The results show that there is a more obvious difference between the instruction features of malicious code and the instruction features of normal code. And in the multi-feature classification experiment, the recognition effect of malicious code recognition by combining features is the best. So the malicious code can be recognized more effectively by combining features.

4. Conclusion

In summary, this paper uses data mining and machine learning methods to classify malicious code from three perspectives: text structure layer, byte layer, and code layer, and compares the classification results of different code levels, which shows that the malicious code detection technology based on data mining and machine learning has better recognition accuracy, sensitivity, and specificity for byte-layer features and combined features, and has better recognition accuracy, sensitivity, and specificity for combined features. The model not only improves the accuracy and precision of recognition but also enhances the recognition ability for combined features.

References

[1] Zhang Xiaokang. Research on malicious code detection technology based on data mining and

ISSN 2616-5775 Vol. 7, Issue 1: 72-77, DOI: 10.25236/AJCIS.2024.070111

machine learning [D]. Anhui:University of Science and Technology of China,2009.
[2] Zhang Zhangwang,Li Yanwei. Android malicious programme detection system based on machine learning algorithm[J]. Computer Application Research,2017,34(6):1774-1777,1782.
[3] Haixin Huang, Lu Zhang,Li Deng. A review of malicious code detection based on data mining[J]. Computer Science,2016,43(7):13-18,56.
[4] Cong Yuandong. Research on Android malicious behaviour detection technology based on hybrid model[D]. Heilongjiang:Harbin Engineering University,2016.
[5] Shi Y. Design and implementation of Trojan horse detection system based on data mining and machine learning[D]. Sichuan:University of Electronic Science and Technology,2014.
[6] Feng Benhui. Research on malicious code detection technology based on data mining and machine learning[D]. Hunan:Central South University,2013.