Virtual Sample Generation for Retraining the Malicious PDF Detection Model

Kang HE$^{1,2}$, Long LIU$^{1,2}$, Dong-zhe LU$^3$ and Yuefei ZHU$^{1,2,*}$

1 State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China
2 Information Engineering University, Zhengzhou, 450001, China

*Corresponding author: yfzhu17@sina.com

Abstract. PDF files are adopted for launching cyberattacks because of their popularity and the increasing number of relative vulnerabilities. Machine learning algorithms are developed to detect the maliciousness of PDF files. As the exploits of new vulnerabilities occur, the assumption that the training data and the test data share the same distribution does not hold and the ability of origin model to detect exploits of new vulnerabilities weakens gradually. In a real environment, it is very difficult to obtain numerous samples of exploits with the same CVE. and the machine learning models are difficult to be improved by retraining. Virtual sample generation could be used to generate sufficient virtual samples by small sample sets to improve the generalization of the existing model. A new VSG algorithm based on prior knowledge is proposed in this paper, which performs better than other VSG algorithms in improving the detection on exploits of new vulnerabilities.

1. Introduction
Portable document format (PDF) is widely used for document exchange for its efficiency and stability, and therefore becomes an important carrier for network attacks. APT attacks targeting governments, organizations, and enterprises consider using phishing emails to launch an initial connection with the target host. Most mail servers prohibit executable files as attachments in emails from downloading to hosts for security reasons, documents such as PDF play an important role in recent network attacks. Ordinary users blindly believe that non-executable files are more secure than executable files, and they lose their alertness over attachments such as PDF documents. But PDF files are as dangerous as executable files and have the ability to carry malicious code. Attackers could use the vulnerabilities of PDF to gain illegal access to the host. Defenders attempt to train machine learning models to detect the existence of malicious PDF files, but it is difficult to obtain a large number of samples that exploiting new-discovered common vulnerabilities to retrain existing models to improve their generality. A new virtual sample generation (VSG) algorithm based on prior knowledge is proposed in this paper.

1.1. Cause of Vulnerability
One of the important reasons for the risk of PDF files is the rich functions that Adobe Reader (the most widely used PDF reader) supports for JavaScript code. This allows PDF files to perform complex tasks such as form validation and mathematical calculations. But it also provides the opportunity for attackers to execute arbitrary code using vulnerabilities in the Adobe JavaScript engine. Figure 1 shows an increase in the number of PDF-related Common Vulnerabilities and Exposures (CVE)
discovered since 2000.

![Figure 1. The number of PDF-related CVE from 2000 to 2018](image)

### 1.2. The difficulties of traditional detection methods

Most of traditional commercial software detection technologies are based on signatures and heuristics which could be easily evaded by minor modifications of existing malicious files. Machine learning algorithms are more and more popular in detecting spam, malware, and network intrusions, and can also be used to detect malicious PDF files[1]. Static and dynamic features are used to train PDF classification models [2], [3]. The difference is that static feature vectors can be obtained directly by parsing documents, while dynamic feature vectors need to be obtained by monitoring the behavior of PDF sample activities in a virtual environment.

Machine learning models have achieved high accuracy on the test dataset, and the accuracy of some algorithms are even as high as 99%. However, on one hand, with the discovery of zero-day vulnerabilities, new attack approaches are developed and the assumption that the training data and test data follow the same distribution does not hold. The generalization ability of machine learning models to detect new exploit decreases. On the other hand, there is a natural race between attackers and defenders in the field of network security[4]. Attackers always attempt to find the weaknesses of the algorithm and modify the relevant features of malicious PDF files to mislead the models to classify them as benign files[5]. Therefore, it is necessary to improve the robustness of the original machine learning model.

An important method to enhance the generalization ability of machine learning models is retraining. The original model will be loaded and trained again with the samples that cannot be correctly classified to improve the model parameters. But a practical problem in the field of network security is that it is very difficult to obtain a large number of new exploit samples of the same CVE. The exploit of new CVE captured by security researchers is usually isolated, which is difficult to improve the original model by simple retraining.

### 1.3. Virtual sample generation methods

VSG is to generate virtual samples by expanding a small sample dataset, and train machine learning model with generated data to reduce overfitting and improve the generalization ability of the model[6]. The main technologies include the following: (1) Generating virtual samples by adding certain regularization constraints to the sample's feature vector space through prior knowledge[7]. Typical examples such as translation, rotation, and symmetric transformation are applied to image samples, which transform the feature values of different positions without changing their meanings. (2) Generating virtual samples based on sample distribution functions, including techniques such as repeated sampling (bootstrap)[7], mega-trend-diffusion (MTD)[8], and diffusion neural network
(DNN)[9]. Bootstrap increases the training weight of new samples, but it does not actually generate new samples. MTD and DNN fill the information in the gap of the samples based on fuzzy theory and information diffusion criteria to obtain new virtual samples. (3) Generating virtual samples by adding small perturbations[10]. Typical examples such as adding Gaussian or uniform distribution perturbations to new sample vector values will generate samples around them[11][12].

In the task of malicious PDF document detection, existing models may be underreported in the face of samples using new zero-day vulnerabilities. To maintain the generalization ability, new samples are used to retrain the original model. Since the rareness of the new samples, it is necessary to generate enough virtual samples. However, most of the samples generated by the above-mentioned VSG technology are limited to the local neighborhood of the seed sample, and the ability to improve model training is limited. This paper proposes a new method for generating virtual samples based on prior knowledge in the field of malicious PDF document detection. It could generate virtual samples with more information with the help of original training data and improve the generalization ability.

2. Algorithm analysis

2.1. Definitions

To illustrate the effectiveness of the prior knowledge-based sample generation method, some definitions are presented first. The theoretical analysis of the relationship between knowledge, model, and data will be performed in the classification task based on machine learning models.

Definition 2.1 dataset $D$, is a set of samples, each sample is represented as $s = (x, y)$ and the dataset $D = (X, Y)$, where $x$ is the feature value vector and $y$ is the class label; $X$ and $Y$ is the feature value vectors and labels of the whole dataset correspondingly.

Definition 2.2 model $M$, is a mapping function from $X$ to $Y$ expressed as $Y = M(X)$, and it outputs the predicted probability value or category label.

Definition 2.3 Knowledge set $K$, is a set of logical propositions expressing knowledge and is represented by $Y = K(X)$. It infers the category label by the calculation of logical propositions.

![Figure 2. The relationship between Data, Model, and Knowledge](image)

The dataset, model, knowledge could all be used to describe data distribution in the classification task. $D$ is sampled on the distribution in the feature value space. $M$ is a probability model established by existing data, and it adjusts the output with the input to match the distribution of the training dataset. $K$ is a rule set generated based on prior knowledge, and classification results can be obtained based on sample characteristics through predicate verification and reasoning. Therefore, $D$, $M$, $K$ could be unified to describe the data distribution and transform to each other. The relationships between them are shown in figure 2. The data can be trained into a model using a machine learning algorithm and the
model could classify the inputs according to the distribution. Knowledge can be abstracted from the existing model and the model can also be designed based on prior knowledge. Knowledge can add constraints to the feature vector values and generate ideal samples, and knowledge can also be inferred from data directly. Since data, model, and knowledge can describe the same distribution, the equivalence needs to be defined first.

Definition 2.4 metrics $p = P_\alpha(x)$, is the evaluation value vector of a model $x$ over dataset $A$. The evaluation metrics include accuracy, recall, F1 score and etc.

Definition 2.5 equivalence. Model $x$ in $C = \{x \mid \|P_\alpha(x) - p_0\| < \epsilon\}$ is equivalent to each other, where $p_0$ is the target performance and $\epsilon$ is the tolerance scope. If $x_1, x_2 \in C$, they are called equivalent over dataset $A$ and represented by $x_1 \cong x_2$.

To determine whether the distributions are equivalent, $M=\text{train}(D)$ and $M=\text{Design}(K)$ are executed first and the evaluation of models is performed. Based on the correspondence between $D, M,$ and $K,$ a method for generating virtual samples based on prior knowledge is proposed. The working scenario of the VSG method is that an effective model has been trained by the existing training dataset, and it shows great performance on the training set and the test set. However, with the discovery of zero-day vulnerabilities and the development of adversarial skills, the generalization ability of the original model has declined.

The defender attempts to enhance the generalization ability of the model by retraining with the new-discovered samples. However, it is very difficult to obtain a large number of samples exploiting the same vulnerability from the wild. At the initial time, only a single zero-day vulnerability sample can be found, so the original model cannot be effectively improved. An important method to cope with the lack of training samples is the VSG technology. Traditional VSG methods include adding noise to samples, and change the weight of new samples do not work well when the number of new samples is too small, and the new sample generated by traditional VSG methods falls in the local neighborhood of the original sample. Therefore, this paper proposes a new VSG method based on prior knowledge based on the specific application scenarios of malicious PDF documents.

### 2.2. Algorithm Description

The idea of the algorithm is shown in figure 3. When the model after retraining by adding a single sample does not work well, the sample will be further analyzed and transformed into knowledge. The knowledge will guide the generation progress of virtual samples for retraining. The details of the algorithm are shown in table 1.

The difference between the newly added sample and the original training set samples on each feature subset will be calculated first. If the newly added sample on the feature subset is similar to part of training samples, these feature subsets are regarded as old knowledge. If the new sample is significantly different from all the samples in the training set, the value vector on these feature subsets is regard as new knowledge and will be fixed. The values of the rest feature sets can be filled by sampling on the training set.

On the one hand, compared with adding Gaussian noise to the samples and changing weights of samples, this generation approach can vary the samples in a larger range. On the other hand, it is implicitly related to the semantic reasoning process of knowledge, which is analogous to the conjunction of knowledge and has certain interpretability. For example, in the detection scenario of a malicious PDF document, a new exploit that a new storage location of malicious code is discovered. The VSG by resampling and weight shifting are not enough, because they ignore the fact that different exploit codes can be stored in the same location for triggering. Our method could migrate the code in training samples to this location to generate more virtual sample sets in a larger range. This semantically corresponds to the expansion content in this location. The sample encoding method, trigger function, and data obfuscation techniques can also be added to the new sample to generate a large number of virtual samples.
Figure 3. The flow chart of the new-proposed VSG

Table 1. The virtual sample generation algorithm

| Input: training dataset $D$, new sample $s$ |
|------------------------------------------|
| 1. $M = \text{train}(D)$ |
| 2. Feature set $F$ is divided to $F_i$ based on realistic functions, $F = \{F_1, F_2, \ldots, F_n\}$ |
| 3. Calculate the distances between $s$ and samples in $D$ |
| 4. Infer knowledge $K(D, s)$ based on the distances |
| 5. Generate virtual samples based on $K(D, s)$: |
| 6. for $x$ in $D$: |
| 7. if label($x$) == label($s$): |
| 8. The value vector of $s$ over feature subsets $F_i$ that has the distances larger than $L$ are preserved and others are sampled in $D$, the generated sample labels are the same as $s$ and $x$. The generated samples are represented as $G_1$ |
| 9. else: |
| 10. The value vectors of $s$ over feature subsets $F_i$ that have the distances larger than $L$ are preserved and others are sampled in $D$, the generated sample labels are the same as $s$. The generated samples are represented as $G_2$ |

Output: $G = G_1 \cup G_2$

### 3. Experiments and results

This paper verifies the proposed VSG algorithm to improve the model's ability to detect a class of samples represented by an isolated sample. The experiment is designed on the task of detecting malicious PDF files. The malicious sample set $Mal$ used for training and testing comes from VirusTotal, which collects a large number of known malicious PDF samples; the benign sample set $Ben$ is collected from the Internet and it has been tested by various anti-virus systems. $Mal$ contains 10965 malicious samples and $Ben$ contains 9,000 benign samples. The collection deadline for these samples is April 2011. To train a machine learning model to detect PDF documents, feature extraction is performed on the PDF files. A 148*150 feature matrix is generated as the feature space of the samples. The convolutional neural network (CNN) algorithm is adopted for detection in this paper and it is famous for its high accuracy. The structure of the CNN model is constructed by one of the most current frameworks – Keras. The experiment is performed on a computer Dell T630, which has a
configuration of 2 Intel Core E5-2630v4 CPUs, 96GB memory and 4 Nvidia GeForce GTX1080 Ti GPUs. The structure of CNN is shown in Figure 4.

The model use 1/3 of the samples in Mal and Ben for training and the rest for testing. The performance of the model is shown in table 2. The accuracy and the recall rate are both over 99%.

|               | Mal   | Ben  |
|---------------|-------|------|
| Classified as malicious | 10497 | 0    |
| Classified as benign    | 4     | 9000 |

|               | Mal   | Ben  |
|---------------|-------|------|
| Precision     | 100%  |      |
| Recall        | 99.96%|      |
| F1            | 99.98%|      |

To test the generation of the CNN model, we collected 25 malicious PDF samples exploiting CVE-2013-0640 and constitute dataset N. The discovery time of these samples is at least two years after the collection deadline of the training data. In the following experiment, only 1 sample will be picked out for generating virtual samples and retraining the CNN model. The accuracy over N is shown in figure 5 and the average performance is shown in table 3. The virtual samples generated by our method performs better than bootstrap and adding Gaussian noises in improving the generalization of the original model in most cases. It could effectively detect the malicious PDF samples exploiting the same CVE with the representing one.
Figure 5. The detection accuracy of improved CNN model over $N$

Table 3. The average detection accuracy of improved CNN model by single sample

| method                  | Bootstrap | Adding Gaussian noise | This paper |
|-------------------------|-----------|-----------------------|------------|
| Average accuracy        | 0.624     | 0.5984                | 0.9728     |

4. Conclusions
A new VSG method based on the transformation between knowledge, data, and model is proposed in this paper. The generated samples could effectively improve the generalization ability on samples exploiting new CVE by various techniques with the help of as less as one representing sample. It also expands the variation range of seed sample compared to traditional VSG techniques.

Acknowledgments
The authors thanks to the education service of VirusTotal and VirusShare for providing malicious PDF samples to train classification models, and the support of the National Key Research and Development Program of China, number 2016YFB0801505.

References
[1] Laskov P and Šrndič N 2011 Static detection of malicious JavaScript-bearing PDF documents Proceedings of the 27th annual computer security applications conference pp 373–382
[2] Stevens D 2011 Malicious PDF documents explained IEEE Security & Privacy 9 80–82
[3] Elingiusti M, Aniello L, Querzoni L and Baldoni R 2018 Malware Detection: A Survey and Taxonomy of Current Techniques Cyber Threat Intelligence (Springer) pp 169–191
[4] Nissim N, Cohen A, Glezer C and Elovici Y 2015 Detection of malicious PDF files and directions for enhancements: A state-of-the art survey Computers & Security 48 246–266
[5] Xu W, Qi Y and Evans D 2016 Automatically evading classifiers Proceedings of the 2016 network and distributed systems symposium vol 10
[6] Yang J, Yu X, Xie Z-Q and Zhang J-P 2011 A novel virtual sample generation method based on Gaussian distribution Knowledge-Based Systems 24 740–748
[7] Chen Z-S, Zhu B, He Y-L and Yu L-A 2017 A PSO based virtual sample generation method for small sample sets: Applications to regression datasets Engineering Applications of Artificial Intelligence 59 236–243
[8] Poggio T and Vetter T 1992 Recognition and structure from one 2D model view: Observations on prototypes, object classes and symmetries (MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB)

[9] Li D-C, Wu C-S, Tsai T-I and Lina Y-S 2007 Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge Computers & Operations Research 34 966–982

[10] Zhu B, Chen Z and Yu L 2016 A novel mega-trend-diffusion for small sample CIESC Journal 67 820–826

[11] Lateh M A, Muda A K, Yusof Z I M, Muda N A and Azmi M S 2017 Handling a Small Dataset Problem in Prediction Model by employ Artificial Data Generation Approach: A Review Journal of Physics: Conference Series vol 892 (IOP Publishing) p 012016

[12] Lu C and Shen H 2018 Virtual sample generation approach for imbalanced classification 2018 9th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP) (IEEE) pp 177–182