L-001

機械学習系マルウェア検知システムへの中毒攻撃データ生成の特徴量空間拡大検討
Feature Space Expansion of Data Generation in Data Poisoning Attack for Machine Learning Based Malware Detection System

蘇 思遠† 長谷川 皓一† 山口 由紀子† 嶋田 創†
Siyuan Su Hirokazu Hasegawa Yukiko Yamaguchi Hajime Shimada

1. はじめに

機械学習や深層学習技術はマルウェアや悪性通信などの情報の検知において多数の応用がある。一方で、機械学習システムは細工された入力に対するロバスト性に欠けており、この脆弱性を悪用し、機械学習システムに攻撃を加えることができることも示されている[1]。そのため、機械学習ベースのマルウェア検知手法に対して、攻撃者が本攻撃用マルウェアの検知率を下げるために、偽学習データを事前にばらまいて、当該機械学習ベースのマルウェア検知機器ベンダ学習用データに混入させ、識別器の出力をミスリードさせる攻撃が考えられる。そのような攻撃は中毒攻撃(Poisoning Attack)と呼ばれる。中毒攻撃は、サポートベクタマシン、ロジスティック回帰、ニューラルネットワークなどの学習モデルに対して可能であると示されている。したがって、このような攻撃を先んじて研究しそれに対応する対策を立てることが重要である。

本研究では、データセットの特徴量空間を中心に、先行研究の問題点を分析し、偽学習データ生成用の特徴量空間の拡大について検討を行う。

2. 先行研究についての検討

機械学習ベースのマルウェア検知システムへの攻撃対策の初期検討として、高木らは機械学習を用いた識別器に対する攻撃の中での中毒攻撃に着目し、ある特定のマルウェアに対する識別精度を下げるような攻撃を検証した[2]。検証では、既存の機械学習系マルウェア検知に向けた攻撃の提案である MalGAN の手法とデータセットをもとに、データセット中の特定のマルウェアのみ検知率を下げる偽学習データの作成を試みた。

MalGAN は 128 個の API 呼び出しの有無を特徴量としたベクトルを学習データとして、検知に関するスコアの差を報酬として強化学習を行う形で、学習用データのベクトルを更新した。評価の結果、偽学習データ生成用特徴量を複数生成できることを確認したが、それを学習した識別器で有意な検知率の差を出すまでには至らなかった。

高木らの研究で有意な差が出なかった理由として、マルウェアサンプルを生成する際に特徴量空間が少ないことは原因の 1 つであると考える。MalGAN で使用する特徴量である 128 個の API 呼び出しの有無を表す 0 と 1 のベクトルの特徴量空間は$2^{128} = 3.4 \times 10^{38}$である。これは偽学習データの生成に対して不十分だと考える。そこで、本研究では FFRI Dataset 2020 を使用して、マルウェアサンプルを生成するための特徴量空間の拡張について検討する。

†名古屋大学 Nagoya University

表 1 FFRI Dataset 2020 に含まれる表層解析結果

表層解析種類	概要
peid	PEiD による表層解析結果
trid	TrID によるファイル種別推定結果
strings	検体中に含まれる文字列情報
lief	LIEF による表層解析結果

3. 特徴量空間拡張

一般的に機械学習のモデルを構築するために、膨大な量のデータを学習させ、そのデータ内の傾向を取得する必要がある。そのため、まずはデータセットから判別に必要な情報を抽出した特徴量という扱いやすいベクトル表現に変換する。特徴量とは判別に必要なデータをコンパクトに表現するベクトルである。たとえば、先行研究ではマルウェアかどうかを判別するために必要な情報はいくつかの API の呼び出しの有無である。しかし、API の呼び出しの有無では、1 つの API 特徴量は“0”と“1”の 2 つの可能性しかなく、特徴量空間が少ない。

そのため、特徴量空間を広げるために、細かな静的解析結果を持ったデータセットである FFRI Dataset 2020 を利用することを考える。本研究では、予備的な検討として、FFRI Dataset 2020 を利用して拡張できる特徴量空間について検討を行った。

3.1 FFRI Dataset 2020 について

FFRI Dataset 2020 は株式会社 FFRI セキュリティがマルウェア対策研究人材育成ワークショップ(MWS)において研究者向けに配布しているデータセットのうちの 1 つである[3]。このデータセットにはマルウェアの表層解析結果が含まれており、FFRI Dataset 2020 では 1 行 1 データの JSONL 形式でマルウェアファイルと良性ファイルそれぞれ 7.5 万件の表層解析結果を提供している。

3.2 FFRI Dataset 2020 の特徴量

表 1 に示すように、FFRI Dataset 2020 は大きく分けて 4 種類の静的解析の結果で構成されている。

この 4 種類の特徴量の中では、“lief”はカスタマイズされた LIEF(Library to Instrument Executable Formats)で PE ファイルをパースした結果、“peid”は PEiD の再実装である pypeid で抽出された表層解析結果、“trid”は TrID によるファイル種別推定結果、“strings”は GNU strings で得られた印刷可能な文字列のうち冒頭から 100 個分である。

3.3 特徴量空間分析

FFRI Dataset 2020 の各静的解析結果に含まれる特徴量のカテゴリ分けとその数を表 2 に示す。特徴量の総数は 2347 個である。lief で抽出された特徴量は 2112 個と大多数を占めるため、さらに細かくカテゴリ分けしてカウントしている。表層解析結果とは別に、file_size も特徴量としてある。

表 2　FFRI Dataset 2020 の特徴量カテゴリと内訳数

特徴量のカテゴリ		内訳数	特徴量のカテゴリ	内訳数
lief	dos_header	31	peid	114
	rich_header	51	trid	20
	header	33	strings	100
	optional_header	53	file_size	1
	data_directories	120		
	sectoins	196		
	relocatons	13		
	tls	36		
	export	155		
	imports	1200		
	resources_tree	5		
	resources_manager	61		
	signatures	1		
	load_configuration	55		
	entry_point	1		
	virtual_size	1		
	小計	2112		

2347 個という特徴量のカテゴリ数だけでも、高木らの評価よりも大きな特徴量空間になりうるが、各特徴量が取りうる値は 2 値ではなく複数値であるため、特徴量空間はより大きなものとなる。たとえば、lief の“dos_header”という特徴量のカテゴリには、表 3 のような 31 個の特徴量がある。そして、FFRI Dataset 2020 の 7.5 万件のマルウェアにおける、dos_header の各特徴量の値のパターン数は表 3 に示す通りである。この 31 個の特徴量のパターン数を乗じて得られる特徴量空間は4.9×10^{59}であり、先行研究の 3.4×10^{38}に比較すると非常に大きなものとなる。

偽学習データを生成するのに、2347 個の特徴量を全て使用するわけではないが、一部分の特徴量のみ使用しても、高木らの研究より特徴量空間は十分に広がるため、効果的な偽学習データの生成に期待が持てる。

4. おわりに

本研究では機械学習系マルウェア検知システムへの中毒攻撃データを生成するために、先行研究を調査して特徴量空間が少ないという問題点を特定した。そして、その問題点を解決するため、FFRI Dataset 2020 を分析して特徴量空間の拡大について検討した。

検討した結果、先行研究で使った特徴量と比べて、特徴量空間がかなり拡大できることがわかった。特徴量空間は、FFRI Dataset 2020 の lief の dos_header カテゴリを使うだけでも、高木らが使った API の呼び出しの有無での3.4×10^{38}から、4.9×10^{59}になり、一部の特徴量のみを使用しても非常に大きく広げることができることが分かった。

表 3　“dos_header”特徴量のパターン数

特徴量	パターン数	特徴量	パターン数
addressof_new_exeheader	69	reserved [0]	149
addressof_relocation_table	125	reserved [1]	126
checksum	39	reserved [2]	105
file_size_in_pages	79	reserved [3]	108
header_size_in_paragraphs	44	reserved2[0]	162
initial_ip	258	reserved2[1]	173
initial_relative_cs	237	reserved2[2]	172
initial_relative_ss	40	reserved2[3]	84
initial_sp	49	reserved2[4]	133
magic	1	reserved2[5]	81
maximum_extra_paragraphs	42	reserved2[6]	334
minimum_extra_paragraphs	47	reserved2[7]	87
numberof_relocation	56	reserved2[8]	121
oem_id	107	reserved2[9]	75
oem_info	127	used_bytes_in_the_last_page	81
overlay_number	117		

一方で、2347 個の特徴量をすべて利用することは生成器や識別器の学習時間や、偽学習データの生成時間が長くなる点で好ましくない。今後は今回の分析結果を利用して、機械学習系マルウェア検知システムや偽学習データ生成に適した特徴量を選択するための分析を進め、中毒攻撃の評価を行う予定である。

謝辞

本研究を進めるにあたり、データセットを整備いただきましたマルウェア対策研究人材育成ワークショップ組織委員会の皆様に深く感謝いたします。

参考文献

[1] Kathrin Grosse et al., “Adversarial Examples for Malware Detection,” In proceedings of 22nd European Symposium on Research in Computer Security, pp. 62-79, Sep. 2017.
[2] 高木聖也ら, “機械学習を用いたマルウェア検知システムに対する強化学習による敵対的サンプル生成の課題,” 信学技報, ICSS2019-62, pp. 13-18, 2019 年 11 月.
[3] 寺田真敏ら, “マルウェア対策のための研究用データセット MWS Datasets ～コミュニティへの貢献とその課題～,” 情処研報, Vol. 2020-IFAT-139, No. 8, pp. 1-6, 2020 年 7 月.