7W-05

Doc2Vec を用いた静的解析によるマルウェア検知

菅野賢輝 † 三浦紘弥 † 三村守 † 田中秀磨 †
防衛大学校情報工学科 †

1. はじめに

近年，未知のマルウェアは日々出現し，ウイルス対策ソフトでの検知は困難となっている．これに対し，動的解析の結果を，機械学習を用いて分類して未知のマルウェアを検知する手法が提案されている[1][2]．しかしながら，動的解析では解析に時間がかかり，大量のトラフィックに対応することは困難である．

本研究では，自然言語処理技術である Doc2Vec 及び教師あり学習モデルである Support Vector Machine (SVM)を用い，実行ファイル内の可読文字列(Strings)から未知のマルウェアを検知する手法を検討する．

2. 関連技術

2.1 Doc2Vec

Word Vector は単語の意味をベクトルで表現する技術である．これを応用した Paragraph Vector では，Word Vector を元にして文章全体を一つのベクトルとして扱い，文章の意味を考慮してベクトルで表現することが可能である．ベクトル化された文章は，意味が似ていればベクトル間の距離が近くなる特徴を示す．

2.2 Support Vector Machine

SVM は教師あり学習モデルの一つであり，マージン最大化という方法で識別平面を決定することで，汎用性に優れた識別が可能である．パターン認識の問題は線形に識別できるものだけではないが，特徴空間を定義することによりその特徴空間上において線形識別を行う事ができるため，非線形であっても活用することが可能となる．

3. 先行研究

先行研究[1]では，マルウェアの動的解析のログから抽出した API 関数名を実行順に抽出した API コール列を，Paragraph Vector を用いてベクトル化してマルウェア亜種を分類している．先行研究[2]では，Word Vector によって API コール列を単語として扱いベクトル化し，マルウェア亜種を分類している．これらの手法では API コール列を用いるため，動的解析が必要である．そのため，大量のトラフィックへの対応は難しいと考えられる．

本研究では大量のトラフィックへの対応をするため，静的解析を用いて未知のマルウェアを検知する手法を検討する．

4. 提案手法

本研究では，Paragraph Vector の実装である Doc2Vec 及び SVM を用い，Strings から未知のマルウェアを検知する手法を提案する．提案手法の概要を図 1 に示す．提案手法では，実行ファイルから可読文字列を抽出し，これを Doc2vec でベクトルに変換する．変換したベクトルに良性あるいは悪性のラベルを付与して分類器を構築し、その分類器を用いて未知のマルウェアを検知する．今回は，Doc2vec および SVM のパラメータについてはデフォルトの値を用いた．

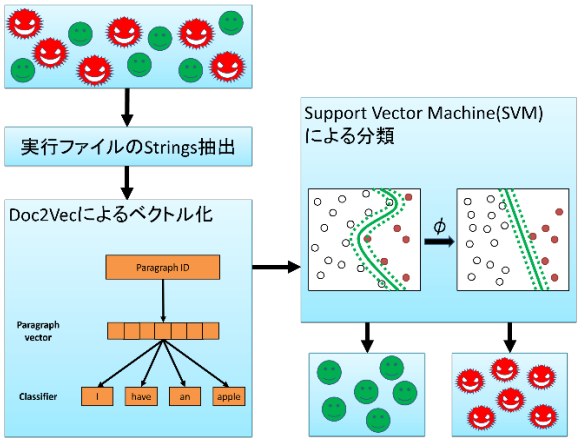

図 1 提案手法の概要

5. 検証実験

5.1 データセット

実験では，Windows10 における Windows フォルダ内にある実行ファイル(2168 個)の可読文字列を，strings コマンドで抽出したものを良性データとして用いる．悪性データは，株式会社 FFRI

Static Malware Detection using Doc2Vec
†Satoki Kanno, Hiroya Miura, Mamoru Mimura, Hidema Tanaka,
†Department of Computer Science, National Defense Academy

が収集したマルウェアの解析結果である FFRI Dataset [3]を用いる．FFRI Dataset の年代ごとの検体数は，2013 年(2644 個)，2014 年(3003 個)，2015 年(3001 個)，2016 年(3040 個)，2017 年(3080 個)である．

5.2 実験方法

検証実験では，交差検証と時系列分析を実施した．交差検証では，Windows フォルダ内にある実行ファイルと FFRI Dataset 2013 を用い，10 分割交差検証を実施した．時系列分析では，Windows フォルダ内の実行ファイル及び FFRI Dataset 2013 を訓練データとし，FFRI Dataset 2014～2017 を各テストデータとした．なお，Windows フォルダ内の実行ファイルは，訓練データとテストデータで重複がないように半数に分類した．評価指標に関しては，Accuracy，Precision，Recall および F-measure を用いた．

5.3 交差検証

交差検証の結果を表１に示す．モデル生成時間は Doc2vec のモデル生成時間，学習時間は SVM の訓練時間である．

表１　10 分割交差検証の結果

Accuracy	0.97
ファイル読み込み時間	21.7(m)
モデル生成時間	8.2(m)
学習時間	2.0(m)
全体時間	32.1(m)

5.4 時系列分析

時系列分析の結果を表２および図２に示す．モデル生成時間は Doc2vec のモデル生成時間，学習時間は SVM の訓練時間である．

表２　時系列分析の結果

	FFRI Dataset 2014	FFRI Dataset 2015	FFRI Dataset 2016	FFRI Dataset 2017
Accuracy	0.90	0.95	0.93	0.91
Precision	0.97	0.97	0.97	0.97
Recall	0.89	0.95	0.94	0.90
F-measure	0.93	0.96	0.95	0.93
ファイル読み込み時間	36.6m	23.7m	25.0m	39.0m
モデル生成時間	4.8m	4.8m	4.7m	5.0m
学習時間	2.9m	2.6m	2.6m	2.6m
全体時間	44.5m	31.1m	24.9m	46.6m

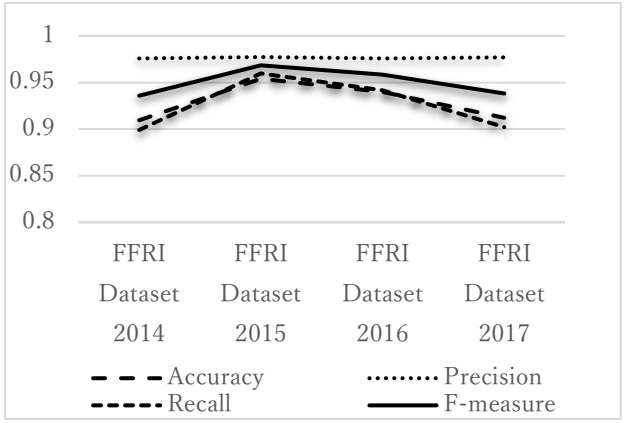

図２　年代毎のグラフ

6. 考察

時系列分析においては，F-measure の値は 0.93 以上となり，４年が経過して精度は大きく下がることはなかった．この結果から，年代にかかわらず，マルウェアには何らかの共通する文字列が含まれている可能性が考えられる．事前に実施可能な処理の時間を除外すると，各年代の約 3000 件の検体の分類に要した時間は１分以内であり，実用的な時間内での検知が可能である．

ただし，今回の実験では良性データとして Windows10 のフォルダ内の実行ファイルのみを用いている．現実には，他にも様々な種類の実行ファイルが存在するため，実環境における精度については検証の余地がある．

7. まとめ

本研究では，Doc2Vec および SVM を用いて実行ファイルの可読文字列から未知のマルウェア検知する手法を提案した．検証実験の結果，F-measure の値は 0.93 以上となり，４年以上が経過しても提案手法が未知のマルウェアに対して有効であることを確認した．実環境における精度の評価や，Doc2Vec モデルのパラメータの最適化については今後の課題である．

参考文献

[1] 佐藤 拓未，後藤 滋樹，武部 嵩礼，「動的解析の Deep Learning による亜種マルウェア推定法」，コンピュータセキュリティシンポジウム 2016 論文集,2016(2)，298-304(2016-10-04)

[2] 佐藤 拓未，後藤 滋樹 ，「Word Vector を用いた亜種マルウェア判別法」， 学位論文,5-35(2017-01-30)

[3] Hatada, M., Akiyama, M., Matsuki, T., Kasama, T., *Empowering Antimalware Research in Japan by Sharing the MWS Datasets*, Journal of Information Processing, Vol.23, No.5, pp.579-588 (2015).