人工知能関連技術のサイバーセキュリティへの活用

乙部　幸一朗

Cylance Japan 株式会社　最高技術責任者

株式会社　技術情報協会
「人工知能の導入による生産性、効率性の向上、新製品開発への活用」
（2018.5.31 発刊書籍収録）

［2］ 人工知能関連技術のサイバーセキュリティへの活用

はじめに

近年，ディープラーニングに代表される機械学習の技術を活用した製品やサービスが多く登場しているが，その多くは一般消費者向けのものであり，またその技術用途も画像認識や言語・音声認識などが主流となっている。企業活動でもデータ解析や業務効率の改善などでも人工知能（AI）関連技術の活用は進みつつあるが，その中でいま注目されているのがサイバーセキュリティ分野への活用である。本稿ではサイバーセキュリティにおける AI 関連技術の活用，とくに機械学習を用いたマルウェアのモデリングについて紹介し，今後期待される技術発展についても解説したい。

1. 背景

企業や組織，そして個人をもターゲットとしたサイバー脅威の数は増加の一途を辿っている。これらは大きくサイバーテロ（Cyber Sabotage），サイバースパイ（Cyber Espionage），サイバー犯罪（Cyber Crime）の 3 つに大きく分けることができるが，大部分がサイバー犯罪と呼ばれる金銭を目的とした活動であり，またサイバー脅威の 9 割以上がマルウェアと呼ばれる悪性プログラムを使用しているとされる。これらのサイバー脅威への対策としては，アンチウイルスと呼ばれるセキュリティソフトウェアが一般的であるが，従来のマルウェア検出技術はシグネチャと呼ばれるプログラムファイルの一部のパターンをブラックリスト化して検出するアプローチが中心となっている。このアプローチでは既出のマルウェアは検出できるものの，未知のマルウェアは検出できないため，攻撃者側は検出を逃れるために常に新しいマルウェアを作成して攻撃に利用するケースが増大しており，従来製品では検出が難しくなっている。

2. サイバーセキュリティと機械学習

そこで最近注目されているのが，サイバーセキュリティへの AI 関連技術の応用だが，その 1 つが機械学習を活用したマルウェアの検知である。具体的には，機械学習を活用して数多くのファイルデータを学習してマルウェアファイルのモデルを作成，そのモデルを元にして新しいファイルに対して推論判定をすることで，未知のマルウェアを含む脅威から防御しようとするという考え方である。

機械学習を元にしたアプローチは，従来のアプローチと比較して基本的な利点がある。まずは，未知のマルウェアを含む高度なサイバー脅威も検出できる可能性が高いこと，そして攻撃者側からすると検出回避がより困難になること，またシグネチャとは違い頻繁なアップデートを要せずとも，汎化能力によってセキュリティ効果が長期間続くことなどである。

2.1 AI の 3 つの波における現状

アメリカ国防高等研究計画局，通称 DARPA（Defense Advanced Research Projects Agency）は，AI 関連技術を 3 つの基本的な種類に分けて定義し，これらを AI の 3 つの波と呼んでいる。

第 1 の波は人手を元にした知識で，人間が特定の機能を実行するために使う規則を定義するというレベルである。コンピューターはこの知識から学習してこれらの規則を自動的に適用し，論理的な推論を行うことができる。しかしながら，この第 1 の波では，より高いレベルの学習に適用することが難しいという課題がある。

第 2 の波は，静的学習である。この波に分類される AI は，機械学習を使用して，するべきことやしてはならないことについての確率的意思決定を行う。この第 2 の波では，システムは学習能力に優れているが，論理的な推論を行う能力に弱点がある。つまり，システムはデータを分類して予測するが，文脈（Context）を理解していないため，あくまで限定的な部分でのみ判断を行うことができる。

次に来るのが文脈適用（Contextual Adaptation）と呼ばれる第 3 の波である。この波では，システムが実世界その

ものの説明モデルを独自に構築する。第 3 の波では，システムは人間とまったく同じように，全体を通して総合的な判断を行うことができ，特徴付けやそのような判断が行われた理由を説明することができる。

実は，機械学習は大量のデータに対する作業を自動化する点に高い能力があることから，サイバーセキュリティでも古くから活用されている。とくにアンチウィルスでは，市場で大多数を占める製品で機械学習の技術が従来から活用されている。しかしながら，そのほとんどは第 1 の波に分類されるもので，つまり人手によって定義された規則やパターンを使用した動作に限定されている。現在注目されているのは DARPA の第 2 の波である静的学習を取り入れているもので，サイバーセキュリティの分野ではこのレベルの技術を実装した製品はまだ少なく，市場の中でもその成熟度はまだ低いのが現状である。

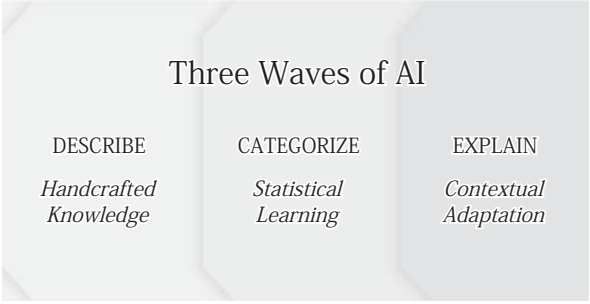

図 1　DARPA が定義した AI 関連技術の 3 つの波

2.2　機械学習の活用

機械学習には数多くの技術手法が存在しているが，それぞれの手法には対象とするデータや学習内容に特徴がある。教師あり学習という前提での手法として，代表的なものとして決定木学習，k 近傍法，サポートベクタマシンなどが挙げられるが，サイバーセキュリティにおけるマルウェアのモデル化手法としては，人工ニューラルネットワーク（図 2）をベースとするものが中心である。その中でも 4 層以上のディープニューラルネットワークを活用した学習が主流となりつつある。サイバーセキュリティの世界ではインターネットを中心として数多くの良質な教師データが存在し，また学習する内容も良性か悪性かというシンプルな判定となるため，機械学習が適用し易い内容であると言えるであろう。

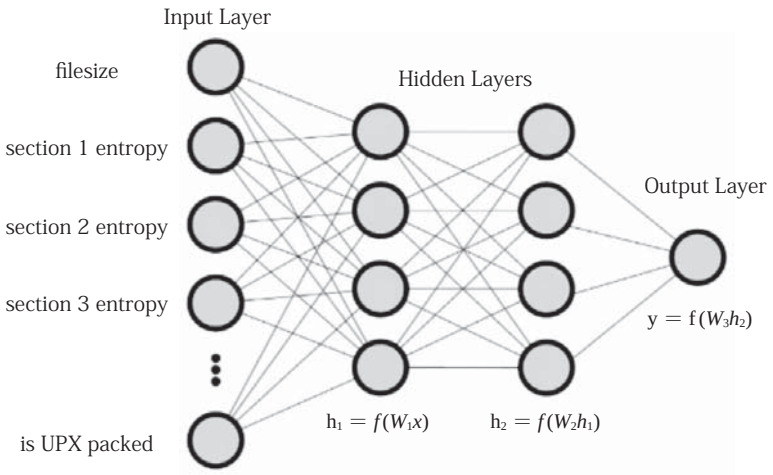

図 2　人工ニューラルネットワークを使ったマルウェアのモデル化のイメージ

3. マルウェアのモデリングの仕組み

ここからは，機械学習によるマルウェアのモデルの概念と仕組みについて解説したい。機械学習を活用したマルウェアのモデリングは，以下の 4 つの基本的なステップから構成される。

①収集：学習のためのデータセットの収集

②抽出：データセットからの特徴量の抽出とベクトル化

③学習：モデルのトレーニング

④検証：未知のサンプルを使ったモデルの検証

ここではモデリングの鍵となる各ステップとその他の技術的な観点について解説する。

3.1 収集

モデルの学習と評価に使用されるデータは，モデルの性能に極めて大きな影響を及ぼす。モデルのトレーニングに使用されるデータが実世界を表現していなければ，モデルが本番環境でうまく動作することはない。マルウェアのモデリングのためには多く PE（実行可能形式）ファイルが必要となり，それぞれのサンプルデータには，トレーニング用の分類子として，良性ファイルか悪性ファイルかといったラベルが必要になる。当然，教師データの肝であるこのラベルデータは慎重に吟味する必要がある。ラベルの付け間違いはラベルノイズとも呼ばれ，モデルの性質を偏らせる可能性があるからだ。高度なモデルを維持するためには，継続的なデータの収集を進めながら，一貫性と正確性を維持するために，ラベル精度のモニタリングも行う必要があるだろう。データセットはさまざまな種類のソースから収集されることが考えられるため，それぞれのソースの信用度や信頼性を正しく評価し，それを考慮に入れた上で活用すべきである。

マルウェアのモデルを作るにあたって，多くのセキュリティ企業に共通する課題としては，アンバランスなデータをどう処理するかということである。一般的にセキュリティ企業はマルウェアのサンプルは多く持っているが，逆にマルウェアでない（良性）ファイルのサンプルが限定的であったりする。この状況は，1 つのラベルが他のラベルより突出して多く存在するという状況を生む。アンバランスなラベルが付いたデータの問題は，さまざまなモデリング戦略によって軽減することができるが，それぞれのラベルを表現する多数のサンプルが存在するのが最も理想的である。多くの特徴はトレーニングセットを使用して生成されるので，特徴セットとデータセットは密接に関連している。データセットは，正規化などの重要な前処理や，単語出現頻度－逆文書頻度（TF-IDF）などの重み付け方式にも影響を与えることになる。

高度なモデルでは，非常に大規模なデータセットが必要となる。データセットが大きければ，より高度なモデルのトレーニングが可能になるが，一方で巨大なデータセットがあれば性能が無条件に良くなるという訳ではない。適切なデータセットには幅広い多様性が必要で，モデルが製品として実装されたときに出会うと考えられる実サンプルを十分に表現しているべきである。

3.2 抽出

用意されたデータセットには，学習に適したデータに変換するための前処理が行われる。機械学習における特徴，すなわち特徴量空間は，データのどの特性を元にモデルを学習するかを正確に指定する。マルウェアのモデリングのためのデータセットとなるのは実行可能形式（PE）ファイルが中心であり，特徴としてはファイルサイズやエントロピーなどの基本的な統計データ，および PE で実質的にプログラムコードが格納されているセクション部分（たとえば，セクションテーブル内の各エントリの名前）に基づく特徴などがまず考えられる。別の派生的な特徴としては，たとえばファイルサイズを底とする 2 の対数なども挙げられるだろう。特徴にはさまざまなものが考えられ，ほかの特徴に基づいて条件的に抽出される特徴や，組み合わせを表現する特徴などがあり，そのため抽出可能な特徴量空間は膨大になる。これは，特徴に適用できる変換が無数にあることを考慮すれば当然のことと言える。

特徴は，学習データをどのように抽出されるかを決定するため，モデルには非常に重要な意味を持つ。どの情報を含

めるかという重要な問題のほかに，情報をエンコードする方法も重要となる。モデルに合った特徴を作成するプロセスは，特徴量抽出と呼ばれる。モデルによっては，特徴のエンコード方法に依存する度合いが他より高いものがある。

直感的には，可能な限り多くの特徴を用意した方が良いと考えがちであるが，使用する特徴が多すぎることには欠点もある。オーバーフィッティング（過学習）が生じるリスクが高くなったり，学習時のリソース消費が増大したりするほか，敵対的な攻撃に対する脆弱性が増すこともあり得るからだ。モデルの能力，解釈可能性，頑健性はすべて特徴によって決まると言っても過言ではないだろう。

特徴が抽出されると，それを元にデータのベクトル化が行われる。特徴に基づいてベクトル化されたファイルデータは数値化され，学習用データとして使用されることになる。

3.3　学習

学習ステップでは，膨大な計算処理が行われることになる。現在主流となっているディープニューラルネットワークは，ハイパフォーマンスなクラウドサーバーの大規模クラスタ環境上でも，トレーニングに数週間から数ヶ月を要することがある。学習が行われた後のモデルによる推論判定は比較的簡単な処理であるが，やはり推論でもメモリや CPU などのリソースを大量に必要とする場合がある。教師あり学習のアプローチでは，分類子のトレーニングを行うためには，入力するデータセットのサンプルに予め適切なラベル（例　良性か悪性か等）が関連付けられていることが重要な鍵となる。

ランタイムは，トレーニングや予測が行われる環境，つまりローカル（例　エンドポイント端末），またはリモート（例　クラウド）を指す。機械学習のランタイムは，新しいサンプルによってモデルを更新できる速度，意思決定の影響，および CPU，メモリ，IO などのリソースに大きく関わってくる。現状では，トレーニングはクラウド上の分散クラスタ環境で行われるのが一般的である。現状では推論もクラウド上で行われるのが一般的であるが，ローカルで実行される例も増えている。クラウドを使わずに，エンドポイント端末上での分散トレーニングを行うような技術アプローチはまだサイバーセキュリティでは見られないが，今後注目したい分野の 1 つである。クラウドを使わないことで，ネットワークトラフィックの削減や機密データの保護など，大きな利点が得られる可能性がある。しかしその反面，エンドポイント間での異なる種類の計算リソースの効率的な活用，不安定な可用性など，多くの課題も予想される。

3.4　検証

学習が終わったモデルは新しいファイルデータに対して推論判定を行うことが可能となる。このステップでは，学習に使用しなかった未知のデータセットを用いて，モデルが正しく推論判定ができるかを検証することになる。もちろん求められる精度の結果が得られない場合には，再び学習を行いモデルの改善を行う必要がある。モデルはブラックボックスになるのが一般的と考えられているが，必ずしもそうではない。人間と対話できるモードをサポートできるモデルには，いくつかの利点がある。モデルはエキスパートからのフィードバックをただちに受け取ることができ，モデルの改良が可能になる。モデルの意思決定がどのように行われているか理解する手段があれば，モデルに対する人間の信頼や信用をより早く確立することができるだろう。モデルを検査して理解するためのツールが実装できれば，トラブルシューティングと診断が可能になる。しかし，こうしたツールは慎重に管理する必要があり，最終的な製品に組み込むことは，知的財産の漏洩や攻撃者に対して脆弱性を露呈するリスクを増やすことにも繋がるため避けるべきであろう。

3.5　モデルの適合度

モデルによっては，実世界をよく表現しているものと，そうでないものが出てくる。モデルが簡素化されすぎていると，新しいデータへの汎化能力があるように見えて，実際には精度が低いという結果が生まれる。このようなモデルは「アンダーフィット（適合不足）」と呼ばれ，モデルに提供される情報の量がモデルの適合能力を超えていて，情報がモデルに完全に取り入れられないという状態を指す。また逆に，学習によりモデルがデータセットを丸暗記した状態になってしまう場合があり，これは「オーバーフィット（過剰適合）」と呼ばれる。過剰適合の場合，モデルはトレーニ

ング対象である特定のサンプルについて学習しすぎて，実世界の新しいサンプルにうまく適応することができなくなる。

次の図では，緑とグレーの点に対する，過剰適合の分類子の決定境界を点線で表している。緑の線は，適切な決定境界を表しており，示された点の分類は完璧ではないものの，新しい点に対する性能は良くなっている。

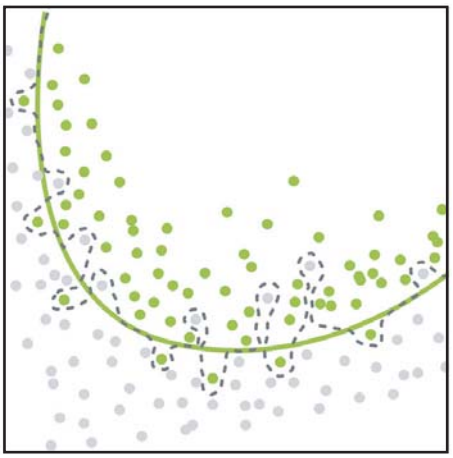

図 3　2 つのクラスからなるデータセットと仮定した分類子から得られる 2 種類の決定境界

適合性の良いモデルは，製品に実装された後もその性能を維持する。これに関連するのがコンセプトドリフトという概念で，データに長期的な非定常の変化があるときに生じるものである。マルウェアのモデルの場合，エンドポイント上で実行可能形式ファイルのデータは毎年変化していくことになるので，対象となるファイルデータの母集団も変化する。モデルはターゲットとする母集団の変化に適応できるように備えておく必要がある。

4.　機械学習活用レベルと今後

機械学習の適用性や効果は幅広い分野で実証されてきたが，サイバーセキュリティへの機械学習の応用は，2013 年頃から始まったばかりで，まだまだ新しい取り組みと言える。サイバーセキュリティの機械学習の活用が始まったといっても，その実装における技術レベルには差がある。この差は下記のようないくつかの主な要素によって把握することができる。これらは，データサイエンスとサイバーセキュリティの両プラットフォームの共通部分を反映しており，モデルの技術レベルを評価するための指標となるだろう。

・ランタイム

学習や予測がどこで行われているか（たとえば，クラウドまたはエンドポイント端末）

・特徴

いくつの特徴が生成されているか，特徴の前処理と評価はどのように行われているか

・データセット

どれだけの量のデータが，どのようなプロセスで選別処理されているか

そしてラベルはどのように生成・提供・検証されているか

・人間との対話

人間はモデルの意思決定をどのように理解し，フィードバックを提供するか

モデルの継続的なモニタリングはどのように行われるか

・適合度

モデルはデータセットをどれだけよく反映しているか，どれほどの頻度で更新する必要があるか

現在のモデル実装の対象はPC端末が中心となっているが，今後はモバイルデバイスやIoTデバイスなど，保護対象となるシステムも多様化していくことになる。またモデルの対象もファイルだけでなく，エンドポイント端末上のさまざまなデータや動作，ネットワーク通信などに広がっていくことが予想される。また，モデルの高度化が進むにつれて，敵対的学習，能動学習，フェデレーションラーニング，モデルの解釈可能性なども重要となり，これらもまた逆に機械学習の最先端研究に深く関わっていくであろう。

サイバー攻撃からの保護と安心できる世界の実現という長い道のりの中で，機械学習を応用する試みは始まったばかりである。機械学習が今後も進化を遂げ，サイバー攻撃を検出して防止するセキュリティ能力の向上に寄与することを期待したい。

文　　献

1）"*2017 Data Breach Investigations Report*", Verizon Enterprise (2017)

2）*John Launchbury*, "*A DARPA Perspective on Artificial Intelligence*", Defense Advanced Research Projects Agency

3）*Cylance Data Science Team*, "*Introduction to Artificial Intelligence for Security Professionals*", Cylance Inc. (2017)