サイバー戦入門　その９

～人工知能の仕組みとその活用～

三村　守（防大情報工学科）

１　はじめに

コンピュータの性能向上やビッグデータの活用が進むのに伴い、様々な分野における人工知能の活用が本格化している。特に、２００６年に提案されたディープラーニング[1]は、画像認識、音声認識、自然言語処理等の分野で顕著な成果をあげており、第３次となる人工知能ブームを引き起こした。人工知能はスマートフォンにおける音声認識、スマートスピーカ、自動運転車等に活用されており、日常生活にも密接に関係するようになってきている。将棋や囲碁のような高度な知性を必要とするゲームにおいても、人工知能はもはや人間を上回っていると言っても過言ではない。人工知能という用語は様々な意味で用いられているが、その中核となっているのは機械学習と呼ばれる人工知能の研究領域における１分野である。機械学習では、人間の学習をコンピュータで再現することを試みる。コンピュータは、正確な記憶、高速な数値計算、繰り返し処理等を得意としており、これらの機械的なタスクでは人間を凌駕している。そのため、単純で規則的なルティーンワークは、コンピュータで容易に自動化することが可能である。また、判断材料を適切に数値化することができれば、人間よりも合理的な意思決定も可能である。これに対し、画像認識、音声認識、言語認識等の高度な認知能力を必要とするタスクは、これまでは人間が得意とされてきた。しかしながら、ディープラーニングの登場に伴い、この一部の分野における人間優位の状況は変化しつつある。ディープラーニングを用いたモデルは従来の機械学習モデルとは異なり、自動的に画像、音声、言語等の特徴を学習し、その特徴を用いて対象の分類を試みる。そのため、従来の機械学習モデルのように、人間が分類のために用いる特徴を指定する必要がなくなった。さらに、人間が気づかない未知の特徴を発見する可能性についても指摘されている。これにより、従来は難しいとされてきた様々なタスクの自動化が可能となり、人工知能が人間よりも高い知性を持つようになるのではないかとの危惧も聞かれる[2]。

機械学習は、サイバーセキュリティの分野においても活用されており、未知のサイバー攻撃の発見等に役立っている。サイバーセキュリティの分野はデジタル

---

[1] Deep Learning 人間の脳を模倣したニューラルネットワークの階層を深めたモデルであり、深層学習とも呼ばれる。

[2] Singularity 人工知能の能力が人間の能力を上回る技術的特異点のことであり、この段階を過ぎると人工知能の能力が爆発的に人間を上回ると考えられている。

技術との親和性が高いため、機械学習技術を適用しやすい。そのため、将来的にサイバー戦は、人工知能ｖｓ人工知能の戦いになるとの見方もある。そのため、サイバー戦の概要を理解するためには、機械学習に関する知識も必要となってきている。

そこで本稿では、人工知能の中核となる機械学習の仕組みについて説明し、サイバーセキュリティ分野におけるその活用例について紹介する。以下、第２節では機械学習を分類し、第３節ではその仕組みについて説明する。第４節ではディープラーニングの概要について触れ、第５節ではサイバーセキュリティ分野におけるその活用例について述べる。

２　機械学習の分類

機械学習とは、人工知能領域における研究分野の１つであり、人間の学習と同等の機能をコンピュータで実現しようとする技術のことであり、統計学を基盤としている。学習にはある程度のサンプルデータを必要とし、これを用いて未知のデータを分析する。その処理は大まかに以下の２つのステップに分類される。

ステップ１：サンプルデータから何らかの規則性や特徴を識別する。

ステップ２：特徴を用いて未知のデータの分類や予測を行う。

機械学習では、これらの処理を数値計算によって実現する。代表的な機械学習のアルゴリズムの分類を表１に、その動作の概要を図１に示す。

表１：機械学習の分類

アルゴリズム	概要	例
教師あり学習	入力とそれに対応するラベルを分類器に与えてモデルを訓練し、訓練したモデルで入力に対するラベルを出力する。	回帰分析、サポートベクタマシン、ランダムフォレスト
教師なし学習	入力からデータの規則性を分析し、モデルを構築する。	主成分分析、因子分析、クラスタリング
強化学習	環境から行動を選択し、選択した行動に対して報酬を与えることにより、どの行動を選択すべきかを学習する。	Ｑ学習、Ｄｅｅｐ Ｑ－Ｎｅｔｗｏｒｋ（ＤＱＮ）

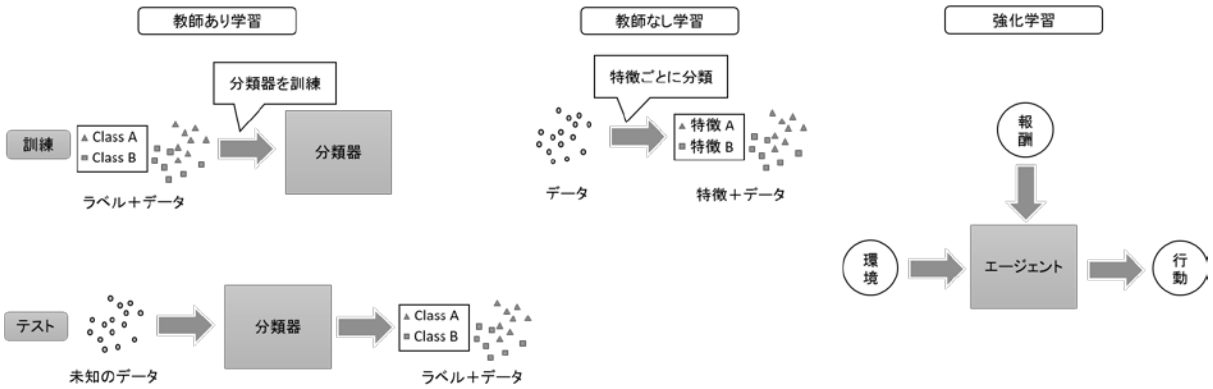

図１：機械学習の動作の概要

もっとも一般的な機械学習のモデルは、教師あり学習モデルである。教師あり学習モデルでは、入力としてサンプルデータとそれに対応する正しい答えを分類器に与えてモデルを訓練し（ステップ１）、その訓練したモデルを用いて未知のデータの答えを予測する（ステップ２）。機械学習の分野では、このサンプルデータに付与する正答のことをラベルと呼ぶ。このように、教師あり学習モデルでは、ラベルが付与されたサンプルデータを必要とし、分類や予測のための事前学習を必要とする。なお、この事前学習は訓練と呼ばれ、訓練に用いない未知のデータの分類をテストあるいは評価と呼ぶ。

これに対し、教師なし学習モデルでは、入力としてサンプルデータのみを与え、そこから何らかの規則性や特徴を識別する（ステップ１）。教師なし学習モデルでは、サンプルデータにラベルが付与されている必要はない。そのため、事前学習を必要としない。

強化学習は、教師あり学習や教師なし学習とはやや異なる考え方に基づいており、サンプルデータの代わりに環境を入力としてエージェントに提供する。エージェントは環境から行動を選択し、選択した行動に対して報酬が与えられる。この動作を繰り返すことにより、そのエージェントは報酬を最大化するように訓練され、どの行動を選択すべきかを学習する。

３　機械学習の仕組み

（１）教師あり学習

一般的な教師あり学習モデルの仕組みを図２に示す。

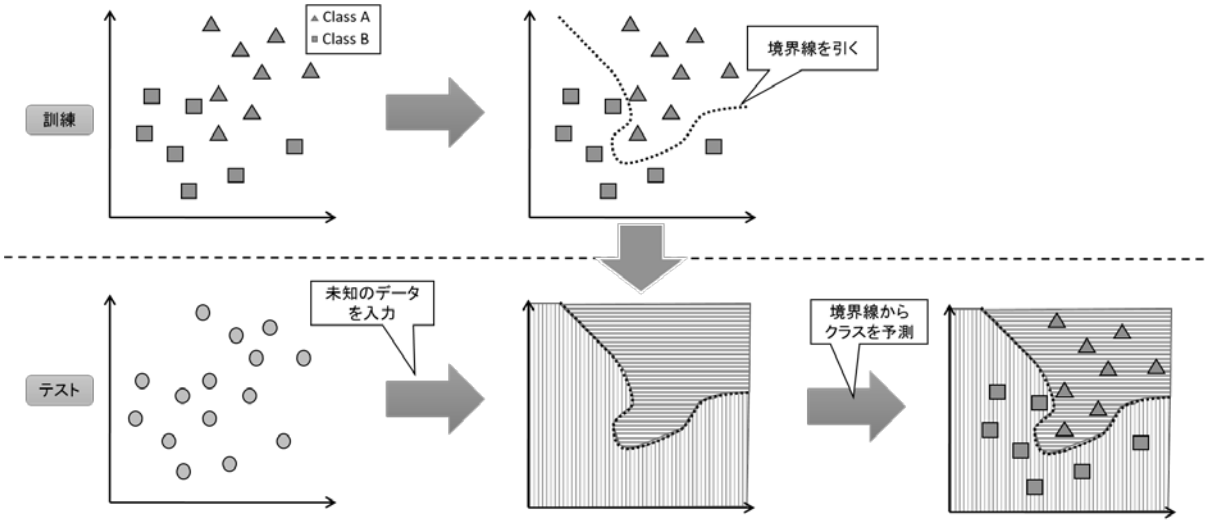

図２：教師あり学習モデルの仕組み

訓練時には、２種類以上のラベルを付与したサンプルを準備する。実際のサンプルは、多次元ベクトルで表現され、３種類以上のクラスに分類することもあるが、ここでは説明のため、視覚化が容易な２次元ベクトルで２つのクラスを用いる。図２の例では、縦横２軸の２次元空間にサンプルがプロットされており、△にはクラスＡ、□にはクラスＢのラベルが付与されている。次に、これらのデータを用いて各クラスを区別するための境界線を引く。この境界線の引き方は、モデルやパラメータによって異なる。

テスト時には、ラベルの付与されていない未知のデータを訓練したモデルに入力し、その予想結果をラベルとして得る。この例では、２次元空間にラベルの付与されていないサンプルが○でプロットされている。これを、訓練時に境界線を引いたモデルに入力し、その境界線との位置関係から△、あるいは□のラベルを予測する。ラベルが△であればそのサンプルはクラスＡ、□であればクラスＢということになる。これにより、未知のサンプルの分類結果を予測することができる。

教師あり学習モデルには、訓練時に１種類のラベルのみを用いる異常検知モデルもある。図３に異常検知モデルの仕組みを示す。

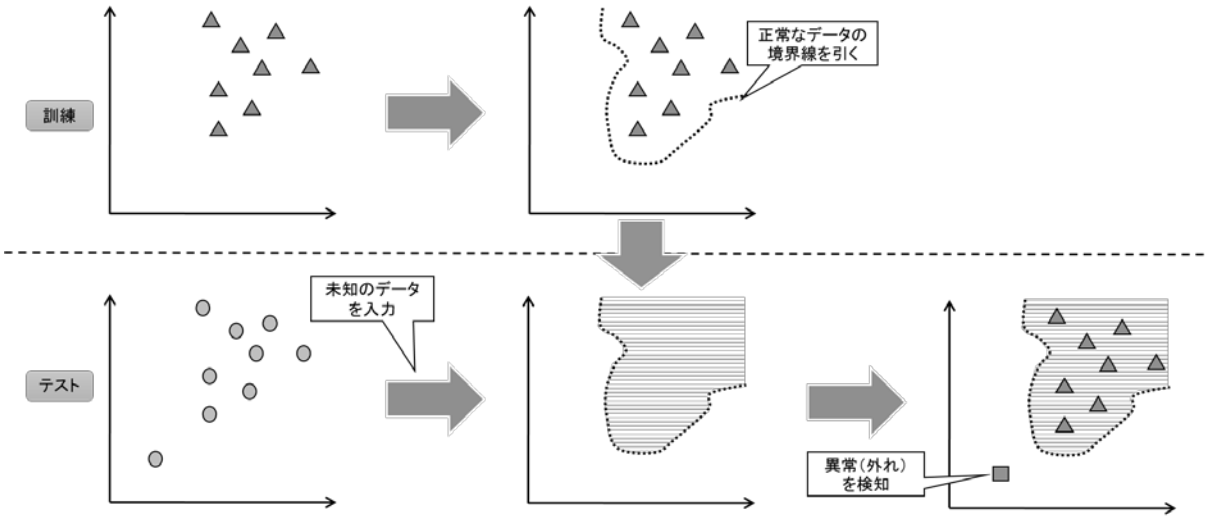

図３：異常（外れ）検知の仕組み

訓練時には、正常のラベルのみが付与されたサンプルを準備する。図３の例では、△の正常のラベルが付与されたサンプルがプロットされている。次に、この正常なデータの境界を定義するための境界線を引く。

テスト時には、ラベルの付与されていない未知のデータを訓練したモデルに入力し、その予想結果をラベルとして得る。異常検知モデルで得られるラベルは、正常あるいは異常のどちらかとなる。この例では、２次元空間にラベルの付与されていないサンプルが○でプロットされている。これを、訓練時に境界線を引いたモデルに入力し、その境界線との位置関係から正常か異常（外れ）かを予測する。ラベルが△であればそのサンプルは正常、□であれば異常ということになる。これにより、未知のサンプルの分類結果を予測することができる。

（２）教師なし学習

一般的な教師なし学習の仕組みを図４に示す。

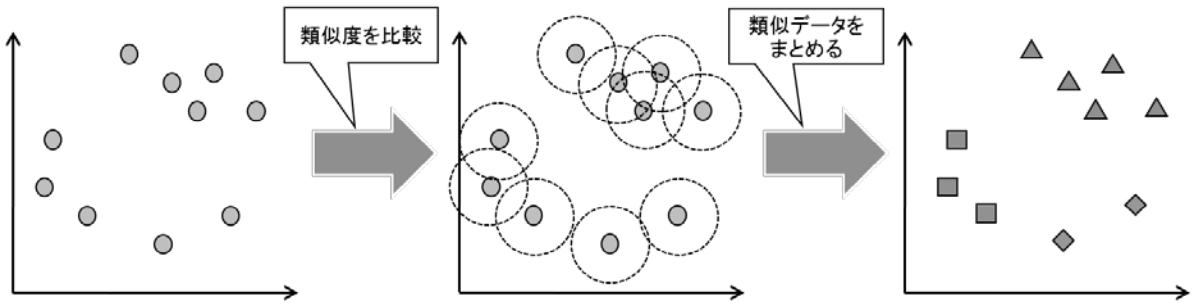

図４：教師なし学習（クラスタリング）の仕組み

ここでは説明のため、２次元空間におけるデータのクラスタリングを例とする。クラスタリングは、類似するデータをグループに分類する手法である。教師なし学習では、事前の訓練やラベルを必要とせず、データのみを準備する。これらのデータに対し、各データ間の類似度を何らかの手法で計算する。この例では、各データから一定の距離の円が描かれており、円が重なるデータを一つのクラスとして△、□、◇の３つのクラスに分類している。この場合、類似度は２点間の距離で計算されており、この距離が一定の範囲内（円が重なる。）の場合には、同一のクラスタに分類されている。つまり、距離の計算手法や一定の範囲を変更することで、様々な手法でデータを分類することができる。

（３）強化学習

一般的な強化学習モデルの仕組みを図５に示す。強化学習では、データを環境として定義し、入力としてエージェントに提供する。エージェントは環境から最適な行動を選択し、その選択した行動によって環境が変化する。ここで、その環境に応じてエージェントに報酬を提供する。エージェントは、この報酬が最大となるように行動を最適化する。例えば、将棋やチェスのようなゲームでは盤面を環境として定義し、次の指し手を行動として定義する。この動作を繰り返すことにより、そのエージェントは報酬を最大化するように訓練され、どの行動を選択すべきかを学習する。

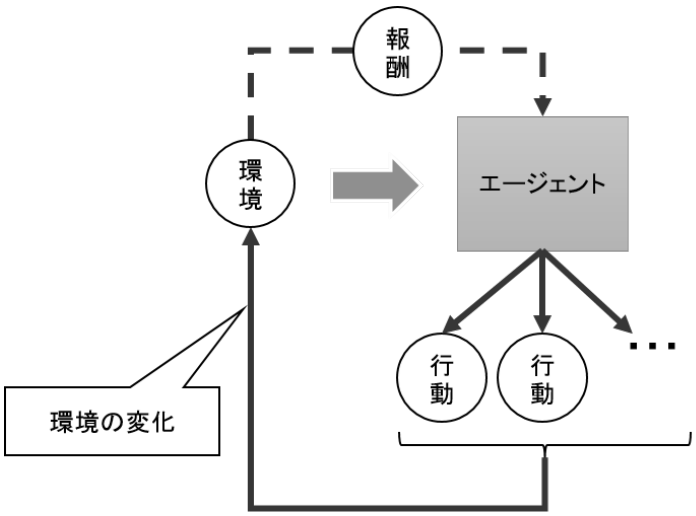

図５：強化学習の仕組み

４　ディープラーニングの概要

（１）ニューラルネットワーク

ディープラーニングの基礎となるニューラルネットワーク[3]は、人間の脳の神経細胞であるニューロンの仕組みを模倣したモデルであり、１９６０年代に提案された。図５に初期のニューラルネットワークである単純パーセプトロンの仕組みを示す。

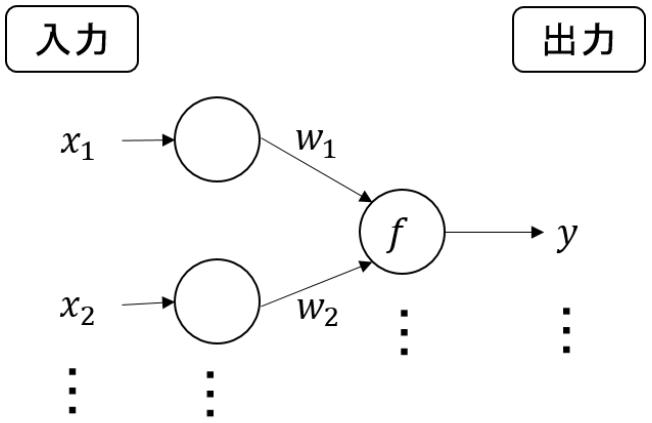

図５：初期のニューラルネットワークの仕組み

この例では、x_1およびx_2を、w_1およびw_2とともにニューロンに入力すると、出力y が得られる。各入力および出力は、０か１のいずれかの値をとる。各入力に対応する重みw は０～１の実数として与えられる。ニューロンでは、各入力とそれに対応する重みをある関数に入力し、その計算結果を出力する。この各入力と重みから０か１の出力を計算する関数は、活性化関数と呼ばれる。

このように、ニューラルネットワークでは各入力の重みによって出力が決定される。つまり、重みを調整することでモデルの出力が変化する。したがって、ラベル付きの多数のサンプルを用いてこの重みを最適化することで、モデルを訓練することができる。しかしながらこの単純なモデルでは、排他的論理和が表現できず、線形分離可能で単純な問題のみにしか適用できないことが指摘されている。

（２）ディープニューラルネットワーク

ディープニューラルネットワーク[4]は、ニューラルネットワークを多層に重ねることで、より複雑な問題を扱うことができるように拡張したモデルである。図６にディープニューラルネットワークの一種である多層パーセプトロンの仕組みを示す。

---

[3] NN (Neural Network)

[4] DNN (Deep Neural Network)

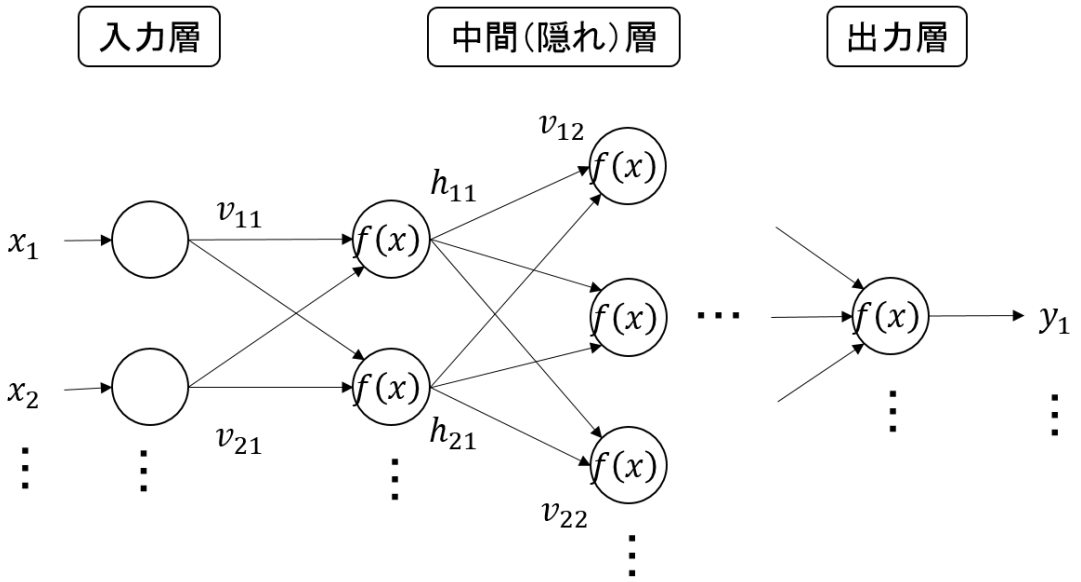

図６：ディープニューラルネットワーク（ＤＮＮ）の仕組み

この例では、x_1およびx_2をモデルに入力すると、出力y_1が得られる。各入力および出力は、０か１のいずれかの値をとる点はニューラルネットワークと同じである。このモデルでは、中間層として複数のニューロンが多層に配置されており、各ニューロンがネットワークで結合されている。入力層に入力された値は、v_{11}およびv_{21}の重みとともに中間層の第１層に入力される。第１層の各ニューロンは活性化関数を経てh_{11}およびh_{21}を出力する。これらの出力は、v_{12}およびv_{22}の重みとともに中間層の第２層に入力される。第２層の各ニューロンは、同様に出力を次の層に入力する。以後、この手順が繰り返され、最終的に出力が得られる。

このように、ディープニューラルネットワークでは各ニューロンの重みとネットワークの構成によって出力が決定される。つまり、各ニューロンにおける重みを調整することで出力が変化する。したがって、ラベル付きの多数のサンプルを用いてこの重みを最適化することで、モデルを訓練することができる。この最適化は、誤差逆伝搬法と呼ばれる手法で繰り返して実施される。この繰り返しの学習の過程において、ラベル毎の特徴は各ニューロンの重みとしてモデルに蓄積される。換言すると、モデルがラベル毎の特徴抽出を自動的に実施していることになる。そのため、ディープニューラルネットワークは教師あり学習モデルとしての特性の他に、何らかの規則性や特徴を識別する教師なし学習の特性も備えていると言える。

５　サイバーセキュリティにおける機械学習

（１）防御技術

これまでに紹介した機械学習のモデルは、様々なサイバーセキュリティの防御技術の研究に適用されている。その主な目的は、未知のサイバー攻撃を検知あるいは分類することである。ウイルス対策ソフトやＩＤＳでは、基本的に定義ファイルやシグネチャに該当する既知のサイバー攻撃しか検知することはできない。しかしながら、近年では未知のサイバー攻撃は増加傾向にある。そこで、機械学習を適用して未知のサイバー攻撃を検知あるいは分類する技術が盛んに研究されるようになった。典型的な機械学習の活用例を図７に示す。

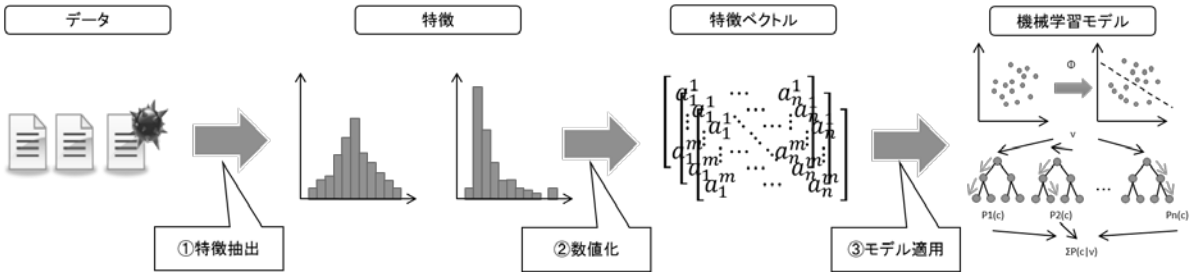

図７：機械学習の活用例

まず、対象とするデータから特徴を抽出する（①）。対象とするデータは、マルウェア、不正コンテンツ、不正通信等の本体あるいはログである場合が多い。主な解析対象のデータを表２に示す。

表２：主な解析対象のデータ

対象	対象	具定例
マルウェア	本体、解析ログ	バイナリファイル、逆アセンブルしたコード、ヘッダ情報、サンドボックスの解析ログ
不正コンテンツ	本体	VBA、JavaScript 等のソースコード
不正通信	通信データ	通信データそのもの、ファイアウォール、ＩＤＳ、プロキシサーバ等のログ

次に、機械学習モデルを適用するため、抽出した特徴を数値に変換する（②）。さらに、各特徴から変換された数値を並べ、特徴ベクトルを作成する。これで、機械学習のモデルを適用することが可能となる。

ここで、未知の攻撃を分類する用途であれば、教師なし学習であるクラスタリング等が適用されることが多い。未知の攻撃を検知する用途であれば、教師

あり学習モデルが適用されることが多い。この場合には、マルウェア等の悪性データの他に、良性データも必要となる。

（２）攻撃技術

機械学習のモデルは、サイバーセキュリティの攻撃技術の研究にも適用されている。

例えば、Ｄｅｅｐ　Ｅｘｐｌｏｉｔ[5]と呼ばれるツールでは、Ｍｅｔａｓｐｌｏｉｔ　Ｆｒａｍｅｗｏｒｋと呼ばれるペネトレーションテスト[6]で使用されるツールと連携し、機械学習を用いて探索から侵入までを自動的に実行することを試みている。Ｄｅｅｐ　Ｅｘｐｌｏｉｔの概要を表３に示す。このツールでは、対象ホストの情報収集、脆弱性の識別、効率的な検査方法の判断、侵入の実行を，強化学習モデルであるＤｅｅｐ　Ｑ－Ｎｅｔｗｏｒｋ（ＤＱＮ）を用いて自動化している。Ｄｅｅｐ　Ｅｘｐｌｏｉｔでは、表３に示す７つの項目から現在の状態を定義し、報酬が最大となるように最適な行動（Ｅｘｐｌｏｉｔモジュールの実行）を選択するように学習する。

表３：Ｄｅｅｐ　Ｅｘｐｌｏｉｔの概要

強化学習のモデル	Ｄｅｅｐ　Ｑ－Ｎｅｔｗｏｒｋ（ＤＱＮ）
状態定義	対象の 0S 対象のポート番号 対象ポートのプロトコル 対象ポートのアプリケーション アプリケーションのバージョン Ｅｘｐｌｏｉｔモジュールの種類 Ｅｘｐｌｏｉｔモジュールの対象
報酬	侵入の成功　＋１ 侵入の失敗　－１

[5] https://www.mbsd.jp/blog/20180228.html

[6] 情報システムの脆弱性の検査

また、敵対的学習[7]と呼ばれる手法では、機械学習を用いた分類器の特徴を逆手に取り、意図的に誤認識をさせて検知を回避する手法も検討されている。図８に敵対的学習の例を示す。

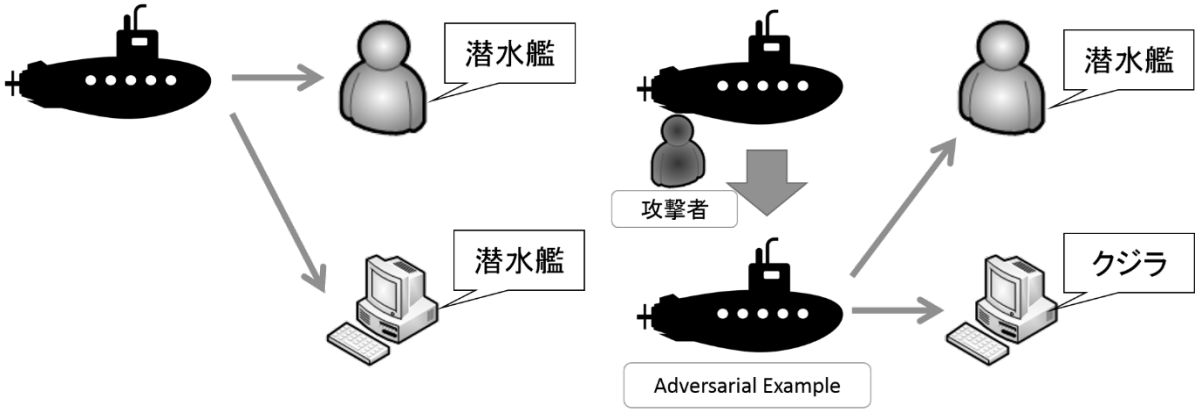

図８：敵対的学習の例

この例では、もとの潜水艦のデータから攻撃者がＡｄｖｅｒｓａｒｉａｌ Ｅｘａｍｐｌｅを作成している。このＡｄｖｅｒｓａｒｉａｌ Ｅｘａｍｐｌｅは、人間からは潜水艦に見えるが、機械学習のモデルは別のデータ（例えばクジラ）と認識してしまう。機械学習の認識の精度は、そのモデルが利用するデータの特徴に依存する。したがって、特徴に利用される部分を類似させ、他の部分はまったく異なるデータ（Ａｄｖｅｒｓａｒｉａｌ Ｅｘａｍｐｌｅ）を用意すれば、誤認識させることが可能となる。特に、ディープニューラルネットワークではこの特徴抽出を自動的に実施するため、モデルがどの特徴を用いているのかが分かりにくい。そのため、このような意図的に誤認識を引き起こす攻撃に気づかない可能性がある点には注意する必要がある。

６　おわりに

本稿では、人工知能の中核となる機械学習の仕組みについて説明し、サイバーセキュリティ分野におけるその活用例について紹介した。第２節では機械学習を分類し、第３節ではその仕組みについて説明した。第４節ではディープラーニングの概要について触れ、第５節ではサイバーセキュリティ分野におけるその活用例について述べた。

---

[7] Adversarial Learning

今後、あらゆる分野において人工知能の活用はますます進むものと予想される。本稿で説明したとおり、その中核となるのは機械学習と呼ばれる研究分野であり、そのモデルの識別能力は特徴の取り方等に依存する。したがって、モデルの構築にあたっては、何を学習の対象とするか、状態をどのように定義するか、あるいは何を報酬として設定するかが重要となる。また、よい特徴を抽出するためには、ラベルが付与された大量のデータが必要となる。したがって、いかに大量のデータを集める仕組みを構築し、どのようにラベルを付与するかも重要な課題である。ディープニューラルネットワークは特徴抽出を自動的に実施することが可能であるが、そのモデルの構築や必要となるデータを収集するのは現在のところは人間である。人工知能ｖｓ人工知能のサイバー戦を制するためには、このような機械学習の特徴を理解し、人間と人工知能の役割分担を適切に実施していく必要がある。