アンラボ・セキュリティレター

Press Ahn

2017.3　Vol.39

MDS に採用された機械学習システム

PRODUCT ISSUE

MDS に採用された機械学習システム

マルウェアが爆発的な勢いで増加している状況で、今のパラダイム構造では人の手でマルウェアを処理することに限界がある。よって機械的に自動処理することが望ましく、セキュリティ業界ではこれを機械学習ソリューションで解決している。アンラボも機械学習システムを知能型脅威対応システム『AhnLab MDS』に採用している。

アンラボでは独自開発した機械学習システムを運営している。これは機械的に作成された規則に従って、機械的な遂行が難しいとされている高水準の分析作業は人が直接行い、再び機械的な部分に関連知識を伝達する柔軟なシステムだ。
アンラボではこのシステムを通じて作成されたエンジンを知能型脅威対応システム「AhnLab MDS エージェント」に適用している。MDSエージェントは PCに潜む疑わしいファイルを検知して、MDS分析サーバーに送信する「疑わしいファイル収集機能」を搭載している。本機能は、感染が疑われる PCを対象にセキュリティ点検を行い、MDSエージェントをインストールする以前に感染されたマルウェアを検知・削除することを可能にする。

[図1] 機械学習システムを採用した知能型脅威対応システム 『AhnLab MDS』

アンラボは疑わしいファイル収集機能をさらに強化するため、機械学習システムを採用するプロジェクトを進めた。
プロジェクトの目標は『スキャン時間短縮』と『検知超過率の最小化』にあった。また、機械学習システムを MDS エージェント以外の応用領域に適用する「拡張性」を検証することも重要なポイントだった。このような前提をもとに複数の技術が検討され、最終的に採用されたのは「機械学習ベースの疑わしいファイル収集」だった。

マルウェア分析とパターン化

アンラボでは、まず任意のファイルを実行せずにそのファイルの『外見』だけで、ファイルがマルウェアであるかどうか判断するという仮説をもとに機械学習プロジェクトを稼動した。

本プロジェクトにおいて「ファイルの実行行為が見えない状態」は、スキャン時間を短縮するために有利な条件と言える。もしファイルの行為をもとに分析するならば、すべての行為領域をモニタリングしなければならないため通常時のシステム性能低下が懸念されるからだ。スピーディーなスキャン実行のためには、情報抽出と規則を素早く適用するように動的情報ではなく『静的情報』のみ読み込むようにした。

通常マルウェアを分析するためには行為情報に注目するが、本システムは複雑かつ収集が難しい行為情報の代わりに『ファイルサイズ』、『ヘッダー』、特定位置の『バイト値』など、ファイルの基本的な「外見(静的分析)パターン」を収集する事にしている。

ファイルの基本パターンを中心に規則を構成すると情報に一貫性が出来上がる。ファイルの行為情報はタイミングや使用環境によって異なる情報を弾き出すことがあるが、ファイルの静的分析資料は時間が経過しても同じ情報を維持するために一貫性を保つことになるのだ。よってファイルの正常/悪性の有無を判別したり、ファイルの実行行為が見えない状態でファイルの基本的な外見パターン(静的分析)を満たすためにどのような方法を使うべきか悩んだ末「機械学習」を採用することになった。

機械学習とパターン化

機械学習は任意のデータから情報を抽出する規則を人の手ではなくコンピューターに算出させることが目標であり、これをシステム的に実現するメリットを持つ。ここでいうシステム的な部分は、分析の専門家らではなく、コンピューターの演算だけで自動化された規則を作り上げることを意味する。

ならばうまく機械学習を遂行するために必要なものは何か。

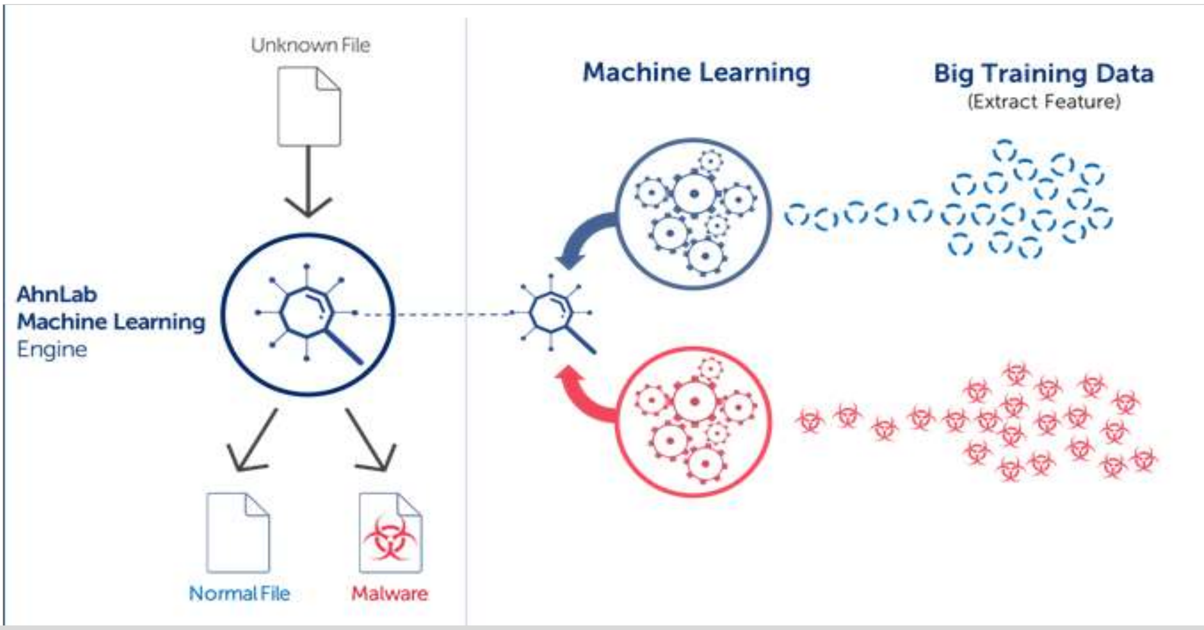
- ✓ 高性能の抽出機 : ファイルの静的データをうまく取り出すための抽出機。アンラボ機械学習システムはファイル当たり155の静的情報を迅速に抽出し、これらの静的情報は常に同じ結果を伝達する。
- ✓ 大量のファイル : 機械学習アルゴリズムは大量のファイルを学習可能でなければならない。仮に、2016年1月の一ヶ月間収集したデータから学習した結果が、2017年にも通用するか考えてみよう。ある程度は通じるが一部は適さないかもしれない。一ヶ月間収集したデータから学習した結果は、その時期の特別なマルウェアトレンドに適した規則で構成されるが、一般的ないくつかのマルウェア分類規則もある程度マッチする。よって例外の時期に収集されたファイルもある程度分類する能力が必要だ。しかしわずか一ヶ月サンプル処理された学習結果がどれほど正確に反映されるだろうか。今PCにある数多くのファイルは2016年1月に作成されたファイルよりも、その前後のタイミングで作成された正常ファイルがほとんどだろう。一ヶ月学習させたモデルでテストしていくら良い結果が得られても、規則を適用した際にほとんどのファイルは今まで学習したことのない時期の正常ファイルがスキャン対象になる。これは誤診の可能性を高める原因の一つになるのだ。

アンラボの機械学習システムは一定期間サンプル処理されたデーターではなく、学習時までに収集された全データーを学習することを目標に、どこにもない『PLANT』アルゴリズムを開発して課題を解決している。現在は13.4億個のファイルをサンプリングなしにすべて学習しており、今後も学習を続けられるように構築された。

機械学習システムの拡張性

ここまでアンラボ機械学習システムの疑わしいファイル収集条件の中で、スキャン時間短縮と診断率改善のための取り組みについて紹介した。次に、機械学習システムの拡張可能性について見てみよう。

最近の主要IT企業は自社の機械学習アルゴリズムをオープンソースで公開したりする。もちろん機械学習の特徴上ソースコードの公開はそう大きい問題ではない。大事なのは機械学習アルゴリズムよりもデータレベルにある。Googleや Facebookでソースコードは公開可能かもしれないが、最も重要なビックデーターに関しては公開することはない。



[図2] アンラボ機械学習システムの概要図

かつてオープン APIで Web Platformを構築しようとした試みのように、今後は自分たちの機械学習プラットホームに取り込もうとする動きなのである。これらを利用すれば手軽に機械学習システムを構築できるが、技術的依存性が発生する副作用もある。いくらハードウェアのスマートフォンをうまく作ってもAndroidというプラットホームに隷属することで発展の制約が生ずるように、持続的な拡張性を持つためにはプラットホームに隷属してはならない。

アンラボの機械学習システムは直接作ったアルゴリズムを使用するため、今後スムーズに技術を拡張する土台を構築しており Windows/Linuxなど多様なOSで動作するように構成された。特に Linuxベースの様々なネットワークアプライアンスから簡単に適用できる。Pythonからも適用が難しくないためマルウェアアナリストも簡単にアクセスし、一時システムからも容易に移植できる。本システムは Google tensorflowのように汎用的な機械学習システムではなく、マルウェア分類のための機械学習を目標にしているが、PLANTアルゴリズムは汎用性を持つため拡張可能な構造となっている。アンラボは MDSだけでなく自社の多様な製品群に機械学習システムを適用する予定だ。

http://jp.ahnlab.com/site/main.do

http://global.ahnlab.com/site/main.do

http://www.ahnlab.com/kr/site/main.do

アンラボとは

株式会社アンラボは、業界をリードする情報セキュリティソリューションの開発会社です。
1995年から弊社では情報セキュリティ分野におけるイノベーターとして最先端技術と高品質のサービスをご提供できるように努力を傾けてまいりました。今後もお客様のビジネス継続性をお守りし、安心できるIT環境づくりに貢献しながらセキュリティ業界の先駆者になれるよう邁進してまいります。
アンラボはデスクトップおよびサーバー、携帯電話、オンライントランザクション、ネットワークアプライアンスなど多岐にわたる総合セキュリティ製品のラインナップを揃えております。どの製品も世界トップクラスのセキュリティレベルを誇り、グローバル向けコンサルタントサービスを含む包括的なセキュリティサービスをお届け致します。

AhnLab

〒108-0014 東京都港区芝4丁目13- 2 田町フロントビル3階 | TEL: 03-6453-8315 (代)