(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6742398号
(P6742398)
(45)発行日　令和2年8月19日(2020.8.19)　(24)登録日　令和2年7月30日(2020.7.30)

(51)Int.Cl.　　F I
G06N 20/00 (2019.01)　　G06N 20/00 130
G06F 21/56 (2013.01)　　G06F 21/56

請求項の数 16　(全 14 頁)

(21)出願番号　特願2018-504758(P2018-504758)
(86)(22)出願日　平成28年6月8日(2016.6.8)
(65)公表番号　特表2018-526732(P2018-526732A)
(43)公表日　平成30年9月13日(2018.9.13)
(86)国際出願番号　PCT/US2016/036408
(87)国際公開番号　WO2017/023416
(87)国際公開日　平成29年2月9日(2017.2.9)
審査請求日　平成30年4月12日(2018.4.12)
(31)優先権主張番号　62/199,390
(32)優先日　平成27年7月31日(2015.7.31)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　517177246
ブルヴェクター，インコーポレーテッド
ＢＬＵＶＥＣＴＯＲ，　ＩＮＣ．
アメリカ合衆国、ブイエー　２２２０３、バージニア、アーリントン、ノース　フェアファックス　ドライブ　４５０１、スイート　１２０
４５０１　Ｎｏｒｔｈ　Ｆａｉｒｆａｘ　Ｄｒｉｖｅ，　Ｓｕｉｔｅ　１２０，　Ａｒｌｉｎｇｔｏｎ，　ＶＡ，　ＶＡ　２２２０３，　Ｕ．Ｓ．Ａ．
(74)代理人　100065248
弁理士　野河　信太郎
(74)代理人　100159385
弁理士　甲斐　伸二

最終頁に続く

(54)【発明の名称】マルウェアの識別とモデルの不均一性のために現場の分類器を再訓練するためのシステム及び方法

(57)【特許請求の範囲】
【請求項１】
マルウェアの識別のために、機械学習分類器を再訓練するための方法であって、前記方法は：
第１機械学習モデルと、複数の第１ファイルと関連する複数の第１特性を示す情報とを受信するステップであって、前記第１機械学習モデルのための訓練データが前記複数の第１ファイルを含むステップと；
前記第１機械学習モデルに基づいて、複数の第２ファイルのための複数のクラス決定を行うステップと；
ひとつ以上の前記複数のクラス決定を判断するステップであって、前記判断することが、前記ひとつ以上の前記複数のクラス決定を確認あるいは修正するユーザー入力を受信することを含むステップと；
前記判断することに基づいて、前記複数の第２ファイルと関連する複数の第２特性を決定するステップと；
前記複数の第１特性の少なくとも一部と、前記複数の第２特性の少なくとも一部とを使って、第２機械学習モデルを決定するステップとを；
備える機械学習分類器の再訓練方法。
【請求項２】
前記第２機械学習モデルを決定するステップは、前記第１機械学習モデルを訓練しテストするために使われた機械学習アルゴリズムを使って、前記第２機械学習モデルを訓練し

テストするステップを含む請求項１に記載の方法。
【請求項３】
前記複数の第１ファイルと前記複数の第２ファイルは、機械実行可能ソフトウェアまたは機械実行可能ソフトウェアによって使用されるファイルタイプを含む請求項１に記載の方法。
【請求項４】
前記複数のクラス決定のうちそれぞれのクラス決定は良性のコンテンツ又は悪意のあるコンテンツのいずれか少なくともひとつである請求項１に記載の方法。
【請求項５】
前記第２機械学習モデルは、企業内のひとつ以上のコンピュータデバイスに分散されている請求項１に記載の方法。
【請求項６】
前記第２機械学習モデルを決定するステップは、前記ひとつ以上の複数のクラス決定の最小値を判断することに基づいて引き起こされる請求項１に記載の方法。
【請求項７】
第１特徴ベクトル表現が、前記複数の第１特性と関連し、第２特徴ベクトル表現が、前記複数の第２特性と関連する請求項1記載の方法。
【請求項８】
前記複数の第１特性の少なくとも一部と、前記複数の第２特性の少なくとも一部とを使うことは、多数の前記複数の第１特性を同数の前記複数の第２特性と交換すること又は前記複数の第２特性を前記複数の第１特性のサブセットに追加することを含む請求項１に記載の方法。
【請求項９】
前記複数の第１特性の少なくとも一部と、前記複数の第２特性の少なくとも一部とを使うことは、前記複数の第２特性を前記複数の第１特性に追加することを含む請求項１に記載の方法。
【請求項１０】
前記判断するステップは、訂正された前記複数の分類を確認するステップと、訂正されていない前記複数の分類を調整するステップとを含む請求項１に記載の方法。
【請求項１１】
前記第２機械学習モデルを決定するステップは、前記複数の分類の分類閾値数を判断することに基づいて引き起こされる請求項１に記載の方法。
【請求項１２】
少なくともひとつのコンピュータデバイスから、複数の第３ファイルに関連する複数の第３特性を示す第２情報を受信するステップを更に含み、第３機械学習モデルが、前記複数の第３ファイルを使って訓練され、前記第２機械学習モデルを決定するステップが、前記複数の第３特性の少なくとも一部を更に使う、請求項１に記載の方法。
【請求項１３】
前記第２機械学習モデルに基づいて、少なくともひとつのファイルが悪意のあるコンテンツを含むことを決定するステップを更に含む請求項１に記載の方法。
【請求項１４】
前記複数の第２ファイルが組織に特有である請求項１に記載の方法。
【請求項１５】
請求項１から１４のいずれか１つに記載の方法をプロセッサによって実行するために遂行される際、コンピュータ読み取り可能な命令を記憶するコンピュータ読み取り可能な記憶媒体。
【請求項１６】
請求項１から１４のいずれか１つに記載の方法を実行するように構成された少なくともひとつのプロセッサとメモリーを含む装置。
【発明の詳細な説明】

【背景技術】
【０００１】
背景
機械学習は、現代のコンピュータの高速処理のパワーを利用してアルゴリズムを実行し、データの挙動や特性の予測を学習する技術である。機械学習技術は、悪意のあるか又は良性の挙動を示すことが知られている１組のファイルのような、公知のクラス（class）や標識（label）によって、１組の訓練（training）サンプル（訓練セット）上でアルゴリズムを実行して、未知のファイルが悪意のあるものか又は良性であるかどうかのような、未知のものの挙動や特性を予想するという特徴を学習する。
【０００２】
機械学習に対する多くの現代のアプローチは、静的な訓練セットを必要とするアルゴリズムを利用する。静的な訓練セット（決定ツリーに基づくもののような）を必要とするアルゴリズムを利用するこのような機械学習アプローチでは、全ての訓練サンプルが、訓練時間には利用できるものであると仮定する。モデルを夫々の新しいサンプル上で更新するというオンライン又は連続的な学習アルゴリズムとして知られる、教師あり機械学習アルゴリズムの分類（class）というものが存在する。しかし、これらのアルゴリズムでは、夫々の新しいサンプルが専門家のユーザによって分類されることを仮定する。
【０００３】
関連する機械学習方法は、バッチモードアクテイブ学習（ＢＭＡＬ）である。ＢＭＡＬは、随時繰り返されるプロセスにおいて、新しいサンプルのバッチに基づき再訓練される新しい分類器（classifier）を構成する。しかし、ＢＭＡＬは、判断ためにユーザに対して、非標識のサンプルを選択することにフォーカスを当てる。ＢＭＡＬは、何らかの客観的な性能基準が一致するまで学習を繰り返し実行する。付け加えると、ＢＭＡＬは、新しいサンプルが追加されたユーザに対し、元の訓練及びテストデータが送られなければならない複数の場所の間に、訓練データが分割されている場合をカバーすることはできない。
【０００４】
他の関連する従来技術の方法は、以下の特許及び公開出願に記載されている。例えば、米国特許第６、５１３、０２５号（以下、‘０２５特許と称す）、タイトル“多段階機械学習プロセス”は、時間間隔による訓練セットの分割と複数の分類器の生成（それぞれの間隔に対して１つの）に関する。時間間隔は、好ましい実施形態においては、周期的/定期的（固定の頻度）である。‘０２５特許は、信頼性モデル（どの分類器モデルを利用するかをシステム入力に基づいて選択する方法）をてこ入れして、どの分類器を利用するかを決定する。更に、この特許における分類器の更新と訓練サンプルの追加方法は、連続的である。更に、’０２５特許では、通信ネットワーク回線に限定されている。
【０００５】
米国特許付与前出願公開Ｎｏ．２０１５００６７８５７号（以下、‘８５７公開と称す）は、“現場の訓練可能侵入検出システム”を指向する。’８５７公開に記載されたシステムは、半教師あり学習（何らかの非標識のサンプルを利用する）に基づいている。この学習は、ファイルではないネットワークトラフィックパターン（ネットフローか他のフローのメタデータ）に基づく。‘８５７公開では、ラプラシアン正規化された最小二乗学習機を利用するが、ユーザに、複数の分類器の間で選択すること、または、複数の分類器の性能の分析を見ることを許す方法は含まない。’８５７公開では、更に、現場（in-situ）のサンプル（クライアント企業からのサンプル）だけを利用する。
【０００６】
米国特許付与前出願公開Ｎｏ．２０１００２９３１１７号（以下、‘１１７公開と称す）、タイトル“バッチモードアクテイブ学習を促進する方法とシステム”には、訓練セットに夫々のサンプルを含めることにより得る“報酬”の推定値（性能向上の推定値）に基づき、訓練セットに含めるべきドキュメントを選択する方法を開示している。この報酬は、未標識のドキュメントまたはドキュメントの長さに関連する不確実性に基づくことができる。’１１７公開には、悪意のあるソフトウエアやファイルを検出することは開示し

ていない。
米国特許付与前出願公開Ｎｏ．２０１２０３１０８６４号（以下、‘８６４公開と称す）、“分類器を進化させるための適応バッチモードアクテイブ学習”は、この技術が画像、音響及びテキストデータ（二値のファイルではなく、及びマルウエア（malware：有害ソフトウエア）の検出のためでもない）に適用することにフォーカスを当てる。更に、’８６４公開は、性能の所定のレベルに典型的に基づく停止基準を定義することを要する。重要なことには、‘８６４公開の方法は、完全なサンプル等を維持する代わりに、そのコーパス（corpus）を特徴ベクトルとして表す部分的な訓練コーパスを与える潜在的な必要性のような、現場（in-situ）の学習を受け入れることができないことである。
【０００７】
現存する機械学習技術は、学習アルゴリズムおよびプロセスを開示しているが、元の訓練者にアクセスできないデータに基づき分類器を増強したり再訓練する方法をカバーしてはいない。現存する機械学習技術は、エンドユーザが、機械学習を実施する上で本来責めを負うべき第三者に対し開示することを望まないデータサンプル上の訓練を可能とはしない。
【０００８】
付け加えると、従来のマルウエア（malware）のセンサーには、マルウエアセンサー（アンチウイルス、ＩＤＳ等）の夫々のインスタンス（instance）は、それらの署名やルールセットが最新のものに更新されているとの仮定の下で、同一であるという本質的な問題がある。そのような場合には、サイバー防御センサーの夫々の配置が同一なので、悪い動作主体ないしマルウエアの著者もセンサーを得ているかもしれず、それで、そのマルウエアをテストし、マルウエアが検出しないようにマルウエアを変更しているかもしれない。このことは、このようなすべてのセンサーを脆弱にするであろう。
【発明の概要】
【課題を解決するための手段】
【０００９】
発明の要旨
前述した従来技術の欠点を克服する実施形態が本明細書に説明されている。これらの及び他の利点は、マルウェアの識別とモデルの不均一性のために、バッチ処理し、教師あり（supervised）により、現場（in-situ）の機械学習分類器を再訓練するための方法により提供される。この方法によれば、ある場所で親分類器のモデルを生成し、それを別の場所又は複数の場所にある１つ以上の現場の再訓練システム又は複数のシステムに対して提供し；現場の再訓練システム又は複数のシステムにより評価された複数のサンプルにわたり、前記親分類器のクラス決定を判断し（adjudicate）；現場の再訓練処理を開始するのに必要な判断サンプルの最小値を決定し（determine）；１つまたは複数の現場のシステムからのサンプルを利用して新しい訓練およびテストセットを作成し（create）；現場の訓練とテストセットを表す特徴ベクトルと、親の訓練とテストセットを表す特徴ベクトルとを混合し（blend）；混合された訓練セットにわたり機械学習を実施し（conduct）；混合されたテストセットと追加された非標識のサンプルを利用して、新しい親モデルを評価し；前記親分類器を再訓練された分類器バージョンにより置き換えるかどうかを選択する。
【図面の簡単な説明】
【００１０】
【図１】図１は、マルウェアの識別とモデル不均一性のために現場の分類器を再訓練するための方法１００の例示的な実施形態を示す。
【図２】図２は、マルウェアの識別とモデル不均一性のために現場の分類器を再訓練するためのシステム２００の例示的なアーキテクチャを示す。
【図３】図３は、システム２００の実施形態による例示的なＧＵＩ３００の画面例（screen shot）を示す。
【図４】図４は、システム２００の別の実施形態による例示的なＧＵＩ４００の別の画面

例（screen shot）を示す。
【図５】図５は、ユーザＡが、ユーザＢおよびユーザＣの両方の間で信頼関係があるが、ユーザＢおよびユーザＣが互いに信頼関係に入れていないシナリオを示す。
【発明を実施するための形態】
【００１１】
詳細な説明
マルウェアの識別とモデルの不均一性のために現場の分類器（in-situ classifier）を再訓練するためのシステム及び方法の実施形態が本明細書に記載されている。これらの実施形態は、上述した問題点を克服する。例えば、この実施形態は、ユーザが駆動する現存するモデルの分類予想と現場の再訓練の確認と修正に基づき、現存する機械学習をベースにした分類モデルの増強を与える。本明細書において、“現場（in-situ：その場）”とは、設置（install）された分類器のインスタンス（instance）の物理的な場所において、機械学習を実施するということを意味する。実際、多数のインスタンスを通じて適用された場合に、この実施形態は、夫々のインスタンスがそのインスタンスに固有のモデルを作成することを可能にする。
【００１２】
好ましい実施形態では、未知の/信頼できないソフトウエア又はソフトウエア・アプリケーション・ファイルが良性であるか悪意のあるかどうかを決定するという問題に適用される。この方法によって生成された子分類器は、特有であるばかりでなく、親分類器の統計的性能を維持又は改善する。特に、この実施形態によれば、ソフトウエア分類の偽陽性率を減らすことが実証されている。この実施形態は、現場の訓練セット（in-situ training set）と呼ばれる、元の訓練セットと補足の訓練セットの組み合わせを利用して親分類器を再訓練することを可能とするように設計されている。現場の訓練セットは、ローカルなインスタンスの環境内で発生し、親分類器を構築した相手を含むいかなる他の者との間で共有される可能性のある潜在的に機密であるデータ又は専有データの必要性を排除する。しかし、この実施形態では、ユーザは、他のユーザとの信頼関係を形成することを選択し、潜在的に機密又は専有データの抽象化を使用して現場の訓練データの一部又は全部を、確実に、共有することを選択することができる。
【００１３】
本明細書に記載された実施形態は、従来技術に対して多くの重要な相違点を含む。上記した従来技術とは対照的に、実施形態は、再訓練以前に新しいサンプルの固定していない間隔でバッチ処理を繰り返し必要とするかもしれないが、すべての新しいサンプルが専門家ユーザにより分類されることを仮定せず、そうでなければ、再訓練バッチに含めることが適当だと仮定しない。これらの相違点により、本システムおよび方法の実施形態を、オンライン又は連続的な学習技術の部類（class）とは異ならせる。
【００１４】
加えて、ＢＭＡＬとは対照的に、実施形態によれば、ユーザが、思いのままに、判断すべきサンプルを選択することを可能にする。同様に、実施形態では、ユーザは、客観的な停止基準を使用するというよりは、再訓練のサイクル数を決定する。さらに加えて、実施形態は、新しいサンプルが追加されたユーザに対し、元の訓練及びテストデータが送られなければならない複数の場所の間に、訓練データが分割されている場合をカバーする。
【００１５】
さらに、‘０２５特許とは対照的に、本明細書に記載の実施形態の現場の学習（in-situ learning）は、現在の分類器の完全な置換を含み、入力空間を細分化することなく、より古いモデルを継続して使用することができる。同様に、本明細書で説明される実施形態は、ユーザにより駆動されるバッチ学習である(全てのイベントが追加の学習に含まれるわけではない)。あるいは、この開示の別の態様では、バッチ学習は自動化プロセスによって駆動されてもよい。’８５７公開に反して、本明細書に記載された現場の実施形態は、半教師あり（semi-supervised）であるラプラシアン正規化最小二乗法学習器とは対照的に、完全に、教師あり（supervised）であり得る。教師なし（unsupervised）及び半

教師あり（semi-supervised）の学習をシステムの態様で実施することもできるが、教師あり学習（supervised learning）が好ましく、これは、例えば、教師あり学習は、未知のサンプルの分類決定をもたらす可能性があるからである。さらに、本明細書で説明する実施形態によれば、クライアント企業からのサンプルと、製造者によって提供されるサンプルとの混合物を使用することができる。‘１１７公開とは対照的に、この実施形態によれば、全ての標識されたサンプルを利用する。’８６４公開とは区別されるように、この実施形態は、単純な停止基準（性能が適切であるか否かの決定を行うユーザを有するシングルパス）を有する。この単純な停止基準によれば、非標識データのバッチと残りの非標識データとの間の距離関数の計算を必要とせず、目的関数の評価に基づいて訓練要素のバッチを選択しない。

【００１６】

この実施形態によれば、ユーザが、分類ソフトウェア/ハードウェアのユーザの配置に、機械学習ベースの分類器を現場で再訓練することを可能にする。再訓練は、全体的な分類器の性能（例えば、偽陽性および偽陰性を低減する）の改善を可能にする。現場での保持（in-situ retaining）は、また、分類モデルの特有なバージョンの作成または生成を可能にする。そのバージョンは、そのインスタンスに固有であり、そのユーザに固有のものであってもよい。調整されたモデルを有することにより、ユーザは、マルウェア生産者が、ユーザのネットワークを危うくしようと試みる前に、検出技術に対してマルウェアをテストすることができないことを保証する。さらに、この調整は、専有または機密性のある特定のタイプのマルウェアの方にバイアスされたモデルを作成するために使用されてもよく、それゆえ、親分類器モデルの作成者にとって利用できないものであってもよい。いくつかの実施形態では、サンプル内容を完全に不明瞭にするが、他のものが再訓練のためにサンプルにてこ入れをすることを可能にする、抽象化されたサンプル表現を使用することによって、複数のユーザの間で共有を容易にすることができる。更に、ユーザは、1つの場所で訓練されたモデルを、それらのネットワーク内または信頼されたパートナーの間で他の場所において共有することを選択することができる。更に加えて、またはこれらの代わりに、現場の再訓練（in-situ retraining）の結果として生成されたモデルは、信頼されたパートナーにエキスポートされるかまたはそれからインポートされることができる。

【００１７】

図１を参照すると、マルウェアの識別とモデル不均一性のために現場の分類器を再訓練するための方法１００の例示的な実施形態が示されている。図示のように、方法１００の実施形態は、１４個のステップのプロセスに関して説明されている。この方法１００は、図２に示すようなソフトウェア/ハードウェアのアーキテクチャで実施することができる。この実施形態では、現場の再訓練プロセス（in-situ retraining process）は、図１において"第三者設備"及び"ユーザ(現場)設備"として示した、２つの物理的に分離された場所に関して行われる。第三者（例えば、マルウェア検出ハードウェア/ソフトウェアを販売する企業)は、基本分類器として知られている分類器の初期バージョン(ブロック１-５を参照)を構築する。この基本分類器は、決定ツリー、サポートベクトルマシン、k-最近接近傍、人工ニューラルネットワーク、ベイジアン（Ｂａｙｅｓｉａｎ）ネットワーク等などの教師あり機械学習アルゴリズムを用いて構築される。第三者は、親訓練およびテストセットを構築する：ブロック１。学習が、同一の種類のサンプルで構成され、全ての所望のクラスを網羅する訓練セットにわたって行われる。この実施形態では、２つのクラス（class）のみが使用され、悪意であるか良性であるかである。サンプルは、コンピュータ実行可能プログラムファイル（ＰＥ３２、ＰＥ３２＋、ＥＬＦ、ＤＭＧ等）と、共通のコンピュータソフトウェア(Ｍｉｃｒｏｓｏｆｔ　Ｗｏｒｄ、Ｍｉｃｒｏｓｏｆｔ　Ｅｘｃｅｌ、ＰＤＦ等)によって使用されるファイルとを含む。第三者は、特徴（例えば、悪意のある及び/又は良性のファイルに存在する可能性の高い特徴)を、訓練セット（例えば、抽出された特徴ベクトルのように)から抽出する：ブロック２。学習を行い、教師あり機械学習アルゴリズムを用いてモデルを作成する：ブロック３。更に、テストセットを用

いてモデルをテストする：ブロック４。このような分類器は、米国特許出願番号１４/０３８、６８２号(ＵＳ２０１４００９００６１号として公開)に記載された方法に従って構築することができる。この出願を参照することにより本明細書に組み入れる。1つまたは複数の分類器を作成して、様々なファイルのタイプをカバーすることができる。

【００１８】

第三者が分類器を作成すると、その分類器は、ユーザ設備（例えば、顧客)に、分類器のインスタンスとして送信/配置される：ブロック５。このような配置５は、複数のユーザ設備（例えば、複数の顧客)、 複数インスタンスの配置の一部であってもよい：。ユーザ設備は、例えば、図２に示すような、システムハードウェアおよびソフトウェアを収容する。本明細書で使用されるように、用語“ユーザ設備”は、企業の物理的ロケーションの一部または全部に配置された、１または複数の現場の再訓練システムを有する複数の物理的ロケーションを含むことができるユーザの企業全体を指す。分類器モデルに加えて、サードパーティはまた、訓練およびテストサンプルから抽出された特徴ベクトルを配信する。特徴ベクトルは、特徴として知られるサンプルの１組の特質または属性に基づくサンプルの要約表現である。特徴ベクトルは、サンプル内容を難読化し、モデル訓練を容易にするサンプルの抽象化された表現である。この特徴には、ファイルヘッダ特性、ファイルの特定の部分または構成要素の存在、n-グラム（n-grams）として知られる連続する２進シーケンス、エントロピー等のような２進表現上の計算のようなものを含むことができる。本明細書に記載された実施形態の重要な特徴は、元の訓練およびテストセットのこの一般化された表現をユーザ設備に送信することである。

【００１９】

引き続き図１を参照すると、元の第三者が作成した基本分類器は、再現可能な現場プロセスにおいて第1の親分類器となる。方法１００は、この基本分類器を使用して、各サンプルについてのクラス(例えば、良性または悪意があるか)を予測するユーザネットワーク上の未知のコンテンツを評価する：ブロック６。一実施形態では、ユーザは、グラフィカルユーザインタフェース（ＧＵＩ）のシステムを使用して、予測されたクラスの一部または全部を検査し、そのサンプルが真に良性かまたは悪意があるかを決定する（例えば、分類を確認または修正する)：ブロック７。この開示の別の態様では、現場の再訓練システム（in-situ retraining system）は、人間の介入なしに、予測されたクラスの一部または全部を検査し、サンプルが良性であるか悪意があるかを決定する（例えば、分類を確認または修正する)。分類を確認または修正する行為を判断（adjudication）と呼ぶ。一実施形態では、再訓練マネージャ（例えば、再訓練マネージャサービスとして例示される)は、ユーザの判断活動を監視し、十分な数の現場のサンプル（in-situ sample）が判断された時を決定する。

【００２０】

この実施形態では、再訓練が起こる前に蓄積されなければならない判断イベントの必要な閾値数が存在する。ユーザが、判断イベントの必要な閾値数を超えると、ユーザは、再訓練を実施することを選択することができる。判断されたサンプルは、１つまたは複数の現場の再訓練システムに保存することができる。信頼関係が存在するという仮定の下で、他のシステムユーザ間と共有することによって、判断サンプルに関する情報は取得することもできる。ユーザが再訓練を開始すると、再訓練マネージャは、判断された現場のサンプルから訓練およびテストセットを作成する：ブロック８。その代替としては、現場の再訓練システムは、人間の介入なしに、再訓練を開始することができる。訓練およびテストセットは、判断されたサンプルのサブセットから選択することができる。再訓練マネージャはまた、再訓練およびテストセットの両方から特徴ベクトルを抽出することができる：ブロック９。次に、方法１００は、これらの現場の特徴ベクトル（in-situ feature vectors）を、親/基本分類器の特徴ベクトル（及び、もしあれば、共有するパートナーからの特徴ベクトル）と混合する（blend）ことができる：ブロック１０。別のモードによれば、一実施形態では、現場のサンプルを追加することなく、親/基本分類器の特徴ベクトル(および共有パートナーからのもの)のサブセットを使用することができる。このサブセッ

トは、利用可能な特徴ベクトルの完全なセットからランダムに選択することができる。1つの形態、加法的方法として知られる混合の実施形態では、現場のサンプルの特徴ベクトルを、親分類器の特徴ベクトルに追加することができる。別の形態、置換方法として知られる第２の混合実施形態では、現場のサンプルの特徴ベクトルは、等しい数の親分類器の特徴ベクトルを置換することができる。別の形態、ハイブリッド法として知られている第３の混合実施形態では、現場のサンプルの特徴ベクトルを、親分類器の特徴ベクトルのサブセットに追加することができる。こうすることで、親セットよりも大きいが、加法的方法によって作成されたものよりも小さい訓練セットを生成することができる。混合にハイブリッド方法を使用することにより、ユーザは、新しい分類モデルに対する現場のサンプルの影響を制限することができる。新しい分類モデルは、親/基本分類器を作成するために使用される同じ機械学習アルゴリズムを使用して、機械学習装置によって訓練される：ブロック１１。新たな分類器が作成されると、それを再訓練テストセットに照らして評価する。この再訓練テストセットは、第三者(基本分類器テストセットの特徴ベクトル)とユーザ設備（再訓練テストセットの特徴ベクトル）の両方からのサンプル特徴ベクトルを含む：ブロック１２。評価１２は、訓練セットに含まれない標識されたサンプルおよび非標識のサンプルの両方に対して生じる。システムＧＵＩは、評価を行う際にユーザを支援するために提供されてもよい。実施形態によれば、どの分類器がよりよいかについて、再訓練マネージャによって提供される自動的推奨を提供することもできる（例えば、図３及び図４を参照)。

【００２１】

引き続き図１を参照すると、評価期間の終了時に、この実施形態では、ユーザは、新しい分類器を受け入れかつ現在の親分類器を置換するか、または新しい分類器を拒否しかつ親分類器を継続するかのいずれかを選択する：ブロック１３。この開示の別の態様では、現場の再訓練システムは、人間の介入なしに、新しい分類器を受け入れかつ現在の親分類器を置換するか、または新しい分類器を拒否しかつ親分類器を継続することができる。いずれの態様の場合も、例えば、ユーザの判断において、または現場の再訓練システムによって、この処理を繰り返してもよい：ブロック１４。新たな現場の分類器が受け入れられると、それは、次のラウンドの現場の再訓練１００のための親/基本分類器となる。ユーザは、さらに、その企業全体の全ての現場の再訓練システムに、再訓練分類器のモデルを配置することを選択することができ、それによって、各システムの親分類器を、新しい再訓練分類器に置き換えることができる。この開示の別の態様では、人間の介入なしに、現場の再訓練システムは、その企業全体の全ての現場の再訓練システムに再訓練分類器モデルを配置し、それによって、各システムの親分類器を新しい再訓練分類器に置き換えることができる。

【００２２】

この実施形態では、連続的な再訓練は、増強（augmentation）のための基礎として、以前のラウンドの訓練およびテストセットを使用する。マルウェアの識別とモデル不均一性のために現場の分類器を再訓練するためのシステムでは、選択的に、元の第三者の基本分類器および関連する訓練およびテストセットに、再訓練を"固定する（anchor）"ことを選択することができる。固定モード（anchor mode）における再訓練時に、元の基本分類器、元の基本分類器の訓練、及び元の基本分類器のテストセット又はそのサブセットは、その後の全ての固定された再訓練のために使用される。

【００２３】

再び図２を参照すると、マルウェアの識別およびモデル不均一性のために現場の分類器を再訓練するためのシステム２００の例示的なアーキテクチャが示されている。システム２００は、ブレードサーバまたはチェーンサーバを含む１つまたは複数のコンピュータサーバによって実施することができる。サーバは、既知、未知および分類されたファイルに関する情報が記憶されているファイルデータベースをホストすることができる。サーバはまた、親モデル、親訓練およびテストセットの特徴ベクトル、現場のモデル、及び現場の訓練およびテストセットの特徴ベクトルをホストすることができる。サービスとして例示

された再訓練マネージャは、サーバ上で実行され、機械学習装置（例えば、機械学習アルゴリズムを実行する機械学習サービス）、及び現場の訓練およびテストセットの特徴ベクトルを使用して、現場のモデルを生成することができる。上述のように、再訓練マネージャは、親モデルおよび親訓練およびテストセットの特徴ベクトルを、現場のモデルおよび現場の訓練およびテストセットの特徴ベクトルに置換することができる。現場のモデルおよび現場の訓練およびテストセットの特徴ベクトルは、新しい親モデルおよび親訓練およびテストセットの特徴ベクトルにそれぞれなる。あるいは、固定された（アンカーされた）再訓練において、元のもの(または固定された訓練が実施される地点に存在するもの)、親モデル、訓練およびテストセットの特徴ベクトルは、現場のモデルおよび現場の訓練およびテストセットの特徴ベクトルと平行なままである。システム２００は、以前に分類されたファイルから特徴を抽出し、現場の特徴ベクトルを作成するために、サーバにおいて、特徴抽出装置(例えば、特徴抽出装置サービス)を使用することができる。特徴抽出装置は、サーバによる入力として受信された未知のファイルから、分類のために特徴を抽出することもできる。サーバは、機械学習装置および現場モデルを使用して、ファイルを分類し、ファイル分類を出力することができる。

【００２４】

システム２００はまた、サーバが出力する現場のモデル、親モデル、テスト結果および分類の表示を、ユーザに提示することを可能にするためのＧＵＩを含むことができる。このＧＵＩはまた、本明細書で説明されるように、例えば、分類を確認または修正し、訓練を選択し、新しい現場のモデル等を受け入れることを選択する等のユーザ入力のエントリーと受理を可能にする。実施形態によれば、サーバは、ＧＵＩを介して入力されたユーザ入力を受理し、本明細書で説明されるようなステップを実行する。この開示の別の態様では、サーバは、現場の再訓練システムによって生成された入力を受理する。

【００２５】

ここで図３を参照すると、システム２００の実施形態による例示的なＧＵＩ３００の画面例（screen shot）が示されている。ＧＵＩ３００は、標識されたサンプル上に現場の再訓練されたモデルの分析を示す、現場の再訓練評価の画面例を示している。図示したように、ＧＵＩ３００は、現場の再訓練されたモデルの分類結果と、基本または親モデルの分類結果との比較を表示することができる。ＧＵＩ３００は、分類スコアの比較を示し、分類スコアは、偽陰性および偽陽性を計数して重みづけする式に基づいてもよく、またはその式から計算されてもよい。ＧＵＩ３００は、ベースモデルのための偽陰性および偽陽性、現場のモデルからの改善と、及び組み合わされたモデル改善(すなわち、現場のモデルとベースモデルを組み合わせたものからの改善)のパーセンテージを示している。

【００２６】

ここで図４を参照すると、システム２００の別の実施形態による例示的なＧＵＩ４００の別の画面例（screen shot）が示されている。ＧＵＩ４００は、非標識されたサンプルの分析を示す現場の再訓練評価の画面例を示している。具体的には、ＧＵＩ４００は、基本モデルに対する新しい現場のモデルを用いて分類の変化を示すグラフを含む。
ＧＵＩ４００はまた、現場の分類器と基本分類器で判断されるように、悪意のある信頼性または可能性が、どれくらいのパーセンテージであるかによって分類されるファイルの数を示す棒グラフを含む(例えば、１８７７は、悪意のある可能性が０％として現場（in-situ）によって分類されたもの)。この棒グラフは、現場の分類器が、悪意のない信頼性が高い（０－１０％)ものか、または悪意のある信頼性が高い（８０－９０％)ものであることを示し、一方、基本分類器は、これらの極端な場合の外側にある信頼度のレベル(例えば、２０－７０％)に分類されたより多くのファイルであって、従って、有用性がより低いファイルを示している。

【００２７】

ここで図５を参照すると、特徴ベクトルを共有するための信頼関係シナリオの図が示されている。この実施形態では、複数のユーザ間の判断されたサンプルの特徴ベクトルの安全な共有が可能である。判断されたサンプルの特徴ベクトルを共有するために、ユーザは、

まず、互いに信頼関係に入ることを選択しなければならない。基本分類器を作成した第三者設備である可能性があるが、必ずしもそうではない、信頼されたブローカーが、元著作者から共有データを受信してレシーバーに転送することにより、サンプルの特徴ベクトルの転送を容易にすることができる。あるいは、信頼関係にある参加者が、ピアツーピア（peer-to-peer）方式で互いに直接データを送信することができる。このデータは、一般に、信頼関係にある参加者間の送信中に暗号化される。図５に示されているのは、ユーザＡは、ユーザＢおよびユーザＣの両方に信頼関係を持っているが、ユーザＢとユーザＣは互いに信頼関係に入っていないシナリオである。このシナリオでは、それ故、ユーザＢは、現場のデータ(特徴ベクトル)を使用することができ、そのデータは、ユーザＡが共有するように選択されるが、ユーザＣが共有するように選択されるデータではない。ユーザＡは、ユーザＢおよびユーザＣの両方からのデータを使用することができる。特徴ベクトルのみを共有することによって、ユーザは、それらが共有する他のユーザからの秘密（confidential）のファイルデータまたは機密（sensitive）のファイルデータを保護することができる。

【００２８】

現場の訓練およびテストセットの構築において共有データが使用される場合、ユーザは、共有された特徴ベクトルの自己および各プロバイダに関して共有データに包含することを優先することを選択することができる。各ソースの優先順位付けは、そのソースの判断されたサンプルから取り出される、訓練およびテストセットのパーセントに変換される。

【００２９】

マルウェアの識別とモデル不均一性のために現場の分類器を再訓練するためのシステムおよび方法の実施形態は、本明細書で説明されるように、先行技術の欠点および不利益の多くを克服する。例えば、本明細書に記載された実施形態は、エンドユーザが、機械学習を実施する上で本来責めを負うべき第三者に対し開示することを望まないデータサンプル上の訓練を可能とすることに挑戦するように対処する。このシナリオの一例は、悪意のあるＰＤＦファイルの識別である。第三者は、分類器を訓練するために、悪意のある及び良性のＰＤＦのコーパスを有することができるが、ユーザのＰＤＦファイルに適用されたときに、分類器は、許容できない数の偽陽性を生成する可能性がある。しかし、ユーザは、ＰＤＦファイルが機密情報又は専有情報を含むことがあるので、不正確にマークされているＰＤＦファイルを共有することを望まない。ユーザが現場で再訓練を行うことを可能にすることにより、ユーザは、そのサンプルを第三者または他のユーザに提供するコストまたはリスクを生ぜずに、訓練セットにそのデータを追加したという利益を得る。この開示の別の態様では、現場の再訓練システムは、そのサンプルを第三者または他のユーザに提供するコストまたはリスクを生ぜずに、訓練セットにそのデータを追加することができる。

【００３０】

加えて、マルウェアの識別とモデルの不均一性のために現場の分類器を再訓練するためのシステムおよび方法の実施形態によれば、マルウェアセンサ(アンチウイルス、ＩＤＳ等)の各インスタンスが同一である(各インスタンスの署名が最新に保たれていると仮定する)場合に、サイバー防御の問題を解決する。サイバー防御センサの各配置が同一であり、悪意のある動作主体またはマルウエアの著者もそのセンサーを得ているかもしれないので、悪意のある動作主体が、マルウエアが検出しないように、そのマルウエアをテストしマルウエアを変更することが可能である。現場の訓練によれば、センサの各インスタンスが、ローカルユーザ以外の誰にも利用可能でないデータ上でそれ自体を調整することを可能にする；この方法は、全ての現場で訓練された分類器モデルが特別であることを効果的に保証する。言い換えると、全てのマルウェア識別モデルのセットは、均一ではなく不均一である。悪意のある動作主体は、もはやそのマルウェアの事前テストに依存することができず、ユーザのコミュニティにわたって発見されるというより大きなリスクを負う。

【００３１】

マルウェアの識別とモデル不均一性のために現場の分類器を再訓練するためのシステム

および方法の実施形態によれば、機械学習の目的のために、潜在的に機密の情報または専有情報の安全な共有の問題にも対処する。ユーザがサンプルの特徴ベクトルを共有するがサンプル自体は共有しないという、ユーザ間の信頼関係を確立することにより、各ユーザは、機密データを露出させることなく、他の作業ができるという利益を得る。

【００３２】

この実施形態は、いくつかの革新的な概念を含む。この実施形態は、各ユーザごとに固有の分類モデルを生成するために、機械学習および現場の再訓練を使用する。本明細書で説明される現場の学習の実施によれば、第三者と現場のデータセットとの組み合わせに基づいて、ユーザが第三者にデータを解放することを必要とせずに、調整するという利点をユーザに可能とする。データセットと、厳密に制御され自動化された機械学習プロセスとの混合によって、ユーザは、不十分な性能をもたらす可能性のある不十分な機械学習法によってもたらされる意図しない誤差を生じにくくなる。このシステムの実施形態によれば、ユーザが、ユーザの優先度を反映しない自動分析に依存するのではなく、再訓練のためにどのサンプルが適格であるかを定義することを可能にする。

【００３３】

この実施形態をテストすれば、広範囲のサンプルのセットについての３０％を超える全体的な偽陽性性能改善により、従来の誤分類された９９％を超える現場のサンプルについての偽陽性率の全ての減少を実証した。これらの改善は、偽陰性率がほとんど増加しないか増加なしで、達成される。さらに、テスト結果は、分類器を再訓練するために異なるデータを使用することが、同じサンプルについて異なる分類挙動の結果をもたらすということも示している。

【００３４】

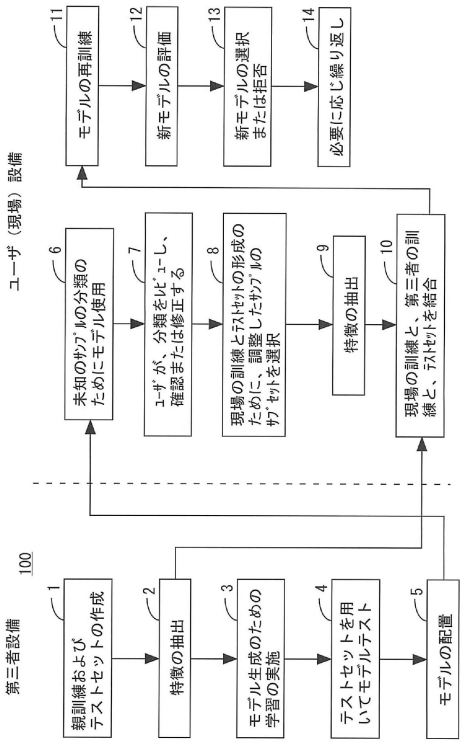
本明細書に記載の実施形態による、現場以前の基本セットの形成を含む現場のプロセスの要旨を、以下に説明する(例えば、ステップ１-５が第三者設備において行われ、ステップ６-１４は、ユーザ設備において行われる)。

1. 基本訓練およびテストセットの作成；
2. 特徴の抽出；
3. モデル作成のための学習の実施；
4. テストセットを用いてモデルテスト；
5. モデルの配置；
6. 未知のサンプルの分類のためにモデル使用；
7. ユーザまたは現場の再訓練システムは、分類をレビューし、確認または修正する；
8. ソース優先順位付けに基づいて、現場の訓練とテストセットの形成のために、判断したサンプルのサブセットを選択；
9. 特徴の抽出；
10. 現場の訓練と第三者の訓練とテストセットまたはそれらのサブセットを結合；
11. モデルの再訓練；
12. 新モデルの評価；
13. 新モデルの配置または拒否；および
14. 必要に応じて、ステップ６-１４の繰り返し。

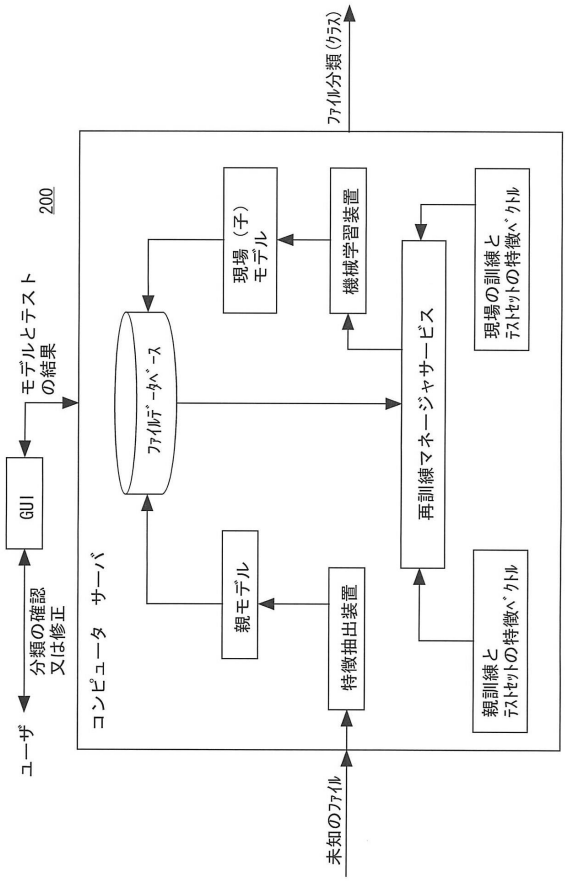
【００３５】

本明細書で使用される用語および説明は、例示だけのために記載されたものであり、限定を意図するものではない。当業者は、以下の特許請求の範囲とこれらと同等のものに定義された本発明の精神および範囲内で、多くの変形が可能であることを認識するであろうし、特に断らない限り、すべての用語が最も広い可能な意味で理解されるべきである。

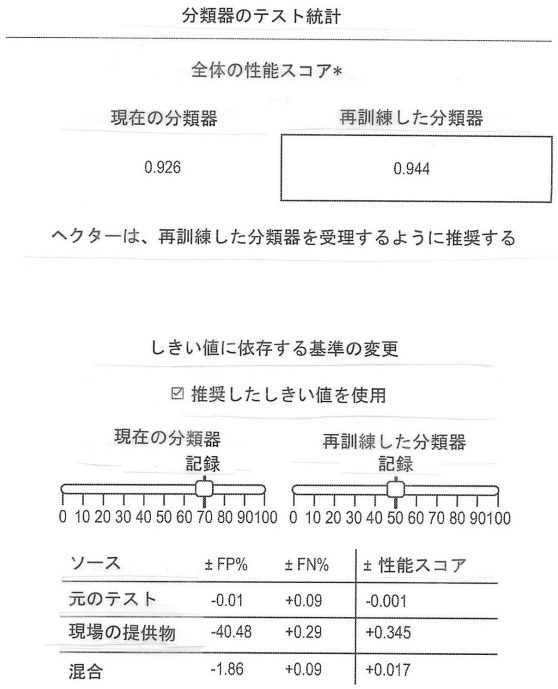
【図１】


第三者設備 100
ユーザ（現場）設備
1 親訓練およびテストセットの作成
2 特徴の抽出
3 モデル生成のための学習の実施
4 テストセットを用いてモデルテスト
5 モデルの配置
6 未知のサンプルの分類のためにモデル使用
7 ユーザが、分類をレビューし、確認または修正する
8 現場の訓練とテストセットの形成のために、調整したサンプルのサブセットを選択
9 特徴の抽出
10 現場の訓練と、第三者の訓練と、テストセットを結合
11 モデルの再訓練
12 新モデルの評価
13 新モデルの選択または拒否
14 必要に応じ繰り返し

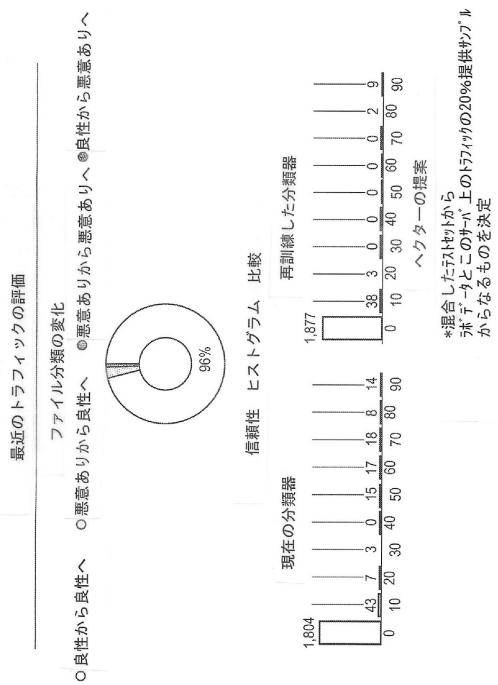
【図２】


ユーザ
GUI
分類の確認又は修正
コンピュータ サーバ
200
モデルとテストの結果
ファイルデータベース
親モデル
特徴抽出装置
現場（子）モデル
機械学習装置
再訓練マネージャサービス
親訓練とテストセットの特徴ベクトル
現場の訓練とテストセットの特徴ベクトル
ファイル分類（クラス）
未知のファイル

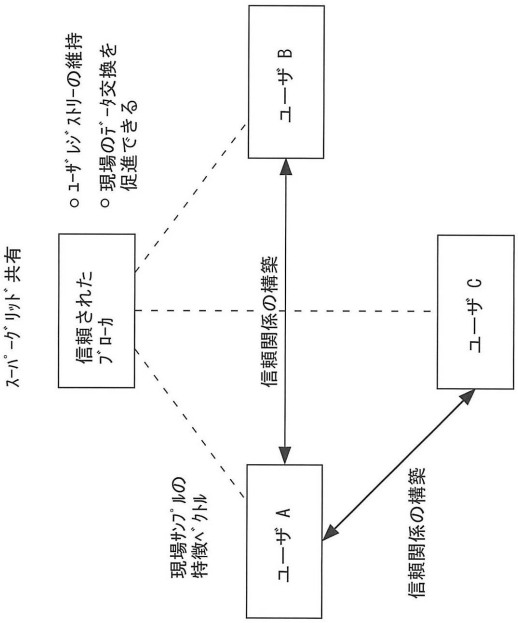
【図３】


分類器のテスト統計
全体の性能スコア*
現在の分類器
再訓練した分類器
0.926
0.944
ヘクターは、再訓練した分類器を受理するように推奨する
しきい値に依存する基準の変更
☑ 推奨したしきい値を使用
現在の分類器
記録
再訓練した分類器
記録
0 10 20 30 40 50 60 70 80 90100
0 10 20 30 40 50 60 70 80 90100
ソース ± FP% ± FN% ± 性能スコア
元のテスト -0.01 +0.09 -0.001
現場の提供物 -40.48 +0.29 +0.345
混合 -1.86 +0.09 +0.017

【図４】


最近のトラフィックの評価
ファイル分類の変化
○良性から良性へ ○悪意ありから良性へ ◉悪意ありから悪意ありへ ◉良性から悪意ありへ
96%
信頼性 ヒストグラム 比較
現在の分類器
再訓練した分類器
1,804
1,877
0 10 20 30 40 50 60 70 80 90
ヘクターの提案
*混合したテストセットからデータとこのサーバ上のトラフィックの20%提供サンプルからなるものを決定

【図５】


ｽｰﾊﾟｰｸﾞﾘｯﾄﾞ共有
信頼されたﾌﾞﾛｰｶ
○ ﾕｰｻﾞﾚｼﾞｽﾄﾘｰの維持
○ 現場のﾃﾞｰﾀ交換を促進できる
ユーザ B
信頼関係の構築
ユーザ C
現場ｻﾝﾌﾟﾙの特徴ﾍﾞｸﾄﾙ
ユーザ A
信頼関係の構築

フロントページの続き

(74)代理人 100163407
弁理士 金子 裕輔
(74)代理人 100166936
弁理士 稲本 潔
(74)代理人 100174883
弁理士 冨田 雅己
(72)発明者 ミセレンディノ，スコット，ビー.
アメリカ合衆国、メリーランド ２１２３０、ボルチモア、ハーパー ストリート １４３４
(72)発明者 クライン，ロバート，エイチ.
アメリカ合衆国、メリーランド ２０９１０、シルバー スプリング、パーシング ドライブ ８１５、アパートメント ２４４
(72)発明者 ピーターズ，ライアン，ブイ.
アメリカ合衆国、メリーランド ２１０７５、エルクリッジ、ハースサイド ウェイ ７５００ ユニット ３０２
(72)発明者 カロルマキス，ピーター，イー.
アメリカ合衆国、メリーランド ２１０４４、コロンビア、スワンズフィールド ロード １１０７７

審査官 塚田 肇

(56)参考文献 米国特許出願公開第２０１５／００６７８５７（ＵＳ，Ａ１）
米国特許出願公開第２０１５／０１３５２６２（ＵＳ，Ａ１）
特開２０１５－０７９５０４（ＪＰ，Ａ）
特表２０１４－５０４３９９（ＪＰ，Ａ）
特開２０１２－０２７７１０（ＪＰ，Ａ）
国際公開第２００５／０９１２１４（WO，Ａ１）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ ２０／００
Ｇ０６Ｆ ２１／５６