特　集　「AI セキュリティの研究動向」

機械アンラーニングの研究に関する現状と課題

Status and Issues of Research on Machine Unlearning

張　海波　　九州大学大学院　システム情報科学府
Haibo ZHANG　　Kyushu University, Graduate School of Information Science and Electrical Engineering.
Haibo0105@gmail.com, https://sites.google.com/view/haibozhang/home

櫻井　幸一　　九州大学大学院　システム情報科学研究院
Kouichi SAKURAI　　Kyushu University, Faculty of Information Science and Electrical Engineering.
sakurai@inf.kyushu-u.ac.jp, https://researchmap.jp/KouichiSakurai-KU

Keywords: Machine Learning, GDPR, Right to be Forgotten, Security, Privacy

1. はじめに

アンラーニング（unlearning）とは、人工知能や機械学習に限定せず、(それまで学んだことを)忘れる、あるいは(習慣や考えなどを)捨て去る行為を指す。その目的は、新しい事柄の習得や、より良い方法を学ぶため、と積極的で、単なる再学習とは区別して用いられている[Cambridge22]。日本語では、学習棄却や学びほぐしとも訳されている[人事労務用語 22]。

機械学習の研究分野で、機械アンラーニング(Machine Unlearning)の概念を、明示的に導入したのは 2015 年 IEEE Security & Privacy での Cao と Yao の論文である[Cao15]。この研究の目的は、システムに、記憶しているデータの一部を忘れさせた上で、次の学習訓練を行うことであり、その実現手段として、機械アンラーニングのアルゴリムを提案している。

実社会においては、欧州が自身の域内の個人データ保護を規定する法として、2016 年 4 月に制定した GDPR（General Data Protection Regulation, 一般データ保護規則）を公式に 2018 年 5 月から施行した[GDPR16]。その後、複数のネット系大企業が、数百億円規模の違反金を課せられるニュースが続く。さらには 2022 年 11 月、日本企業の欧州系列会社が取引先の情報漏洩で、数百万円の制裁金を課せられる事件も報道された。機械アンラーニングの研究論文数は、増え続ける一方であるが、その背景には、GDPR を意識した現場の実システムへの応用があり、多くの論文では、研究動機と応用に GDPR を明記している。

GDPR の尊重する個人データの削除権は、情報漏洩だけではなく、人工知能の訓練データに利用されている個人情報や、芸術作品などのオプトアウトや著作権問題とも深く関係していることにも注意する[MIT22]。

国内外の研究動向

機械アンラーニングを始めて論じた Cao と Yao の論文[Cao15]の引用は、Google Scholar で調査すると、12 月末現在で 300 件弱である。発表された国際会議が、Security & Privacy であることからも、これは機械学習システムのプライバシー漏洩対策がその動機である。さらに、発表された IEEE 系の会議は、この方面では最上位にランキング[Zhou22]されていることにも注意する。

海外では、この分野のトップ国際会議 NDSS2023 において、機械アンラーニングアルゴリムに関する深い研究論文も採択された[Warnecke23]。Arxiv で公開された研究論文の多くは、人工知能やセキュリティとプライバシー研究分野でのトップ国際会議で発表されている。さらに、優れた入門[Mercuri22]や解説論文[Nguyen22]も,Arxiv に掲載され始めた。

国内では、ブログで解説記事がある程度が現状であり、ほとんど研究事例がない機械アンラーニングである。唯一、Cao と Yao の研究[Cao2016]を引用している披田野ら［披田野 16]の主題は、攻撃者が悪性データを注入し、予測モデルを操作することで、出力情報のみから、ユーザの個人情報を復元する攻撃に関する研究であり、個人データの削除後の再訓練を扱うアンラーニングとは対照的である。

この解説論文は、著者の研究グループの先行調査[Zhang22a]を発展させた学術論文草案[Zhang22b]を参考に、その後、現在までの最新動向調査も付加したものである。

2. 機械アンラーニングの導入背景

機械学習のセキュリティの観点からは、攻撃者が学習に用いるデータセットに悪意のあるデータを注入することで、機械学習モデルにバックドアを仕掛けることもできる[Liu22]。その結果、攻撃者はモデル内のほとんどのプライベートデータを盗み出すことができるようになる危険性が存在する。

対策としては、そのような悪性データをデータセットから削除した上で再トレーニングすることも必要となる[Baracaldo17]。しかし，単純な再訓練では大量の計算量と時間を消費してしまう。この課題を解決すべく，より効率的な手法として，機械アンラーニングが提案されている[Bourtoule21]。

「アンラーニング」とは、機械学習モデルを再訓練し、データの一部を、あえて利用しない、忘れた状態で新たな予測モデルを生成することを意味する。学習モデルに対してアンラーニングを行うには、大きく 2 つの方法がある。一つは、データを削除した後の新しいデータセットを、最初から再学習する厳密アンラーニングである。もう一つは，機械学習モデルとデータセットを修正し，近似的にアンラーニングを行う方法である。どちらも、最終目標は、できるだけ効率的にアンラーニング法の精度を向上させることである。

3. GDPR の忘れられる権利

近年、ユーザーの個人情報の使用を規制する法律が増加している。一般データ保護規則 (GDPR) の第 17 条「忘れられる権利(right to be forgotten)」は、ユーザーがそのような要求をした場合、データセットからデータの一部を削除することを事業者に要求している[Shelter21,Graves21]。さらに、プライバシー保護の観点から、ユーザーの権利維持も保障するものである[Al-Rubaie19]。

2012 年の最初の草案から 2016 年に正式な欧州連合法になるまで、忘れられる権利は当初、インターネット検索エンジンがユーザーのプライバシーを使用する際に拘束することを意図していた。それがさらに、第 17 条では、ユーザーがあらゆる個人データの削除を要求した場合、検索エンジンは即座に対応を実行しなければならず、拒否することは許されないと、強化されている。

だたし、この法律の施行により、現在話題になっている機械学習技術とユーザー個人情報の過剰利用については、慎重に議論する必要がある。例えば、機械学習基盤システムに対して、学習用に利用した特定の個人データの削除が要求された場合、データを管理する側は、如何に対応すべきか、あるいは対応できるのであろうか。

機械学習の文脈では、忘れられる権利のために、機械学習システムやデータ管理者は、あらゆるデータの削除を希望するデータ所有者からの要求に容易に対応することが求められる[Chen21]。この要求への対応を可能にする手続きこそが、機械アンラーニングである。機械学習アプリケーションは、要求されたデータを学習データセットから削除し、機械学習モデルをゼロから再訓練する、その効率化こそが機械アンラーニングの主題である。

GDPR 第 17 条の出現により、現在も成長を続けインターネットとビッグデータ環境において、ユーザーのプライバシーが悪用され、保護されないという望ましくない現象は制限されるようになる。しかし、プライバシー保護は、データ管理者とデータ保有者の両方の立場から実施されなければならない。忘れられる権利は、データ管理者の視点からは規制されるが、データ保有者の視点からは機能しない。つまり、データ保有者がどのようにプライバシーの侵害に気づくか、どのような場合に個人データの削除を要求するかということは課題である。

4. セキュリティに関する懸念

4.1 機械学習は、どこまで堅固であるか

機械学習モデル自体の脆弱性から、攻撃者は悪意のあるリクエストを多数送ることで機械学習モデルを攻撃し、機械学習サービスを様々な潜在的な危険にさらすことが可能となる[Gao22]。主な危険性を以下に列挙する。

- データプライバシー漏えい(Data Privacy Leakage)：攻撃者は、モデルの脆弱性を利用し、機械学習サービスを呼び出すことで、モデルを学習するためのデータ情報を取得する。
- モデル盗用(Model Theft)：モデル自体の脆弱性により機械学習サービス内のモデルパラメータ情報を攻撃者が推測し、頻繁にサービスを呼び出すことでモデルパラメータ情報を復元する。
- データ汚染(Data Poisoning)：攻撃者は要求プロセスに特定の悪意のあるデータを混ぜることができ、サービスプロセスのフィードバックを通じてモデル学習とその後のモデル推論に影響を与え、モデルを妨害する効果を実現する[Marchant22,Baracaldo18]。

- はぐらかし(Evasion)：攻撃者は、典型的なリクエストに少量のノイズと摂動を加えることで、機械学習サービスに誤った判断を誘導する。

通常、機械学習システムを設計する際、開発者は設計されたシステムが安全で信頼に値することを保証するために、特定の脅威モデルを考慮する。これまでのところ、既存の機械学習モデルの多くは、攻撃者をあまり広く考慮することなく、特定の脆弱な脅威モデルに対してのみ設計・実装されている[Chundawta22]。これらのモデルは、自然な入力に対しては非常に良い性能を発揮するが、現実的な状況では、これらの機械学習モデルは多くの悪意のあるユーザや攻撃者に遭遇する。

Toreini ら[Toreini20]は、社会科学からの信頼に関する考察を、人工知能利用のサービスや製品に提案される信頼性技術に関連付けるための体系的なアプローチを提供している。例えば、攻撃者は、モデルの学習と予測の段階で、悪意を持って入力と出力を変更することができる程度が異なる。また、攻撃者は、何らかの手段でモデルの内部構造にアクセスし、パラメータを盗むことも可能である。

3.2 機械学習における安全性の三大要素

情報セキュリティの３大要素 CIA(Confidentiality, Integrity, Availability)とは、セキュリティシステムやポリシーを設計評価する際の基礎となるな評価基準である。CIA とは、機密性、完全性、可用性のことであり、これら３要素はシステムの脆弱性を特定し、問題に対処し、効果的な解決策を生み出すための方法を明らかにする。

機械学習モデルに対する攻撃も、この機密性、完全性、可用性に影響を与える危険性がある[Surma20]。ここでは、CIA の３要素が機械学習モデルにどのように適用されるかを説明する[Zhang22b]。

- 機密性攻撃とは、機械学習システムが、権限のないユーザが情報にアクセスできないようにしなければならないことを意味する。ほとんどの機械学習プラットフォームは専門的で安全であるが、機械学習モデルプロバイダが提供するアルゴリズムは必ずしも信頼できるものではない [Tramer16]。データ保有者が機械学習サービス MLaaS (Machine Learning as a Service) を利用して予測モデルを学習する場合、攻撃者が意図的に構築した悪意のあるモデルを選択してしまう危険性がある。このようなモデルでは、攻撃者がデータ保有者の貴重なデータをモデルのパラメータに符号化し、最終的にモデルのパラメータを復号することでユーザの個人データを盗み出す[Song17]。
- 機械学習モデルは、学習段階と予測段階の両方で発生する完全性攻撃に対しては、最も脆弱である。攻撃者がモデルを改ざんし、その完全性を破壊した場合、モデルの予測結果は予想から外れることになる。攻撃者は、既存の学習セットを変更したり、悪意のあるデータを挿入追加し、モデルの完全性を損なわせ、予測段階の精度を低下させる [Shen16]。モデルが学習され、予測に使用されるとき、攻撃者は予測されるサンプルにわずかな摂動を加えるだけでよく、それは人間の目には認識できないが、モデルの分類を誤らせるには十分な脅威となりえる。
- 機械学習モデルが利用可能であること自体も攻撃の対象になり得る。例えば、無人運転のシナリオでは、攻撃者は、車両が通過する道路脇に識別が容易なものを作為的に置いた場合、自動運転車を安全保護モードに移行させ、道路脇に停止させることができる可能性がある。

5. プライバシー保護

機械学習におけるプライバシー保護に対する主要なアプローチとしては現在、秘匿計算、モデルプライバシー、連合学習などが提案されている[Al-Rubaie19]。

- 秘匿計算とは、データの送信や計算過程自体を秘密にすることである。現在、秘匿計算を実現するためのアプローチとして、多者間 (Multi-party) 安全計算の理論、準同型暗号モジュールの活用、隔離実行ワードウエア TEE(Trusted Executive Environment)実装などが主流である。秘匿計算は、学習過程におけるデータのプライバシーを保護するために行うことができる。しかし、学習済みモデルがプライベートな学習データの漏洩を引き起こす危険性がある。なぜなら、機械学習モデルはある程度過学習している可能性があるからである：モデル自身が学習データの一部を記憶しているため、公開されたモデルによるプライベートな学習データの漏洩につながる。
- モデルプライバシーは、差分プライベート機械学習と機械アンラーニングのアルゴリズムを含む。差分プライバシーを実現するためによく行われるのは、ノイズを加えることである。ノイズの付加はモデルの性能低下を伴うが、差分プライバシー機械学習の理論では、いかに効果的なノイズを付加するか、いかに最小限のノイズを付加すれば、与えられたプライバシー損失の要件に対して最高の性能を達成できるかを明らかにする。データそのものにノイズを乗せての処理が困難な場合には、ノイズを、学習アルゴリズム自体に適用させる技術[Abadi16]も開発され、そこで差分プライバシーが解析されている。

また、モデルプライバシー研究のもう一つが、本論の主題である機械アンラーニングである。もし、差分プライバシーを実装することが、出力モデルがプライバシー要件を満たすように積極的にアルゴリズムを設計するこ

とと見なされるなら、機械アンラーニングはモデルプライバシーに対する受動的な解決策と言える。これは、ユーザーの「忘れられる権利」を機械学習モデルで実現することを目的としている。
- 連合学習は、データを共有することなく、複数パーティによる連合体機械学習を行うことであり、本質的にはデータアクセスを制限した分散型機械学習フレームワークである。一般的な分散型機械学習と比較すると、連携学習における制約の第一層はデータの分離であり、エンドポイント間でデータを共有せず、バランスを取らず、双方向通信を最小限にとどめる。

6. 機械アンラーニングの特徴と分類

6.1 再トレーニングとの比較

機械アンラーニングの最も直感的な手法は、指定されたデータを削除した後に、学習データセットでモデルを再訓練することである。しかし、この単純な方法では計算コストがかかるため、機械・アンラーニングの第一の目標は計算コストの削減である。1 つの手法は、機械アンラーニングアルゴリズムの結果が再訓練されたモデルと統計的に区別がつかないように、訓練されたモデルを後処理することである[Ginart 2019]。もう一つの手法は、再訓練のコストを削減するために、新しい学習方法を設計することである。例えば、データを異なるブロックに分割し、各ブロックに対して個別のサブモデルを学習し、サブモデルの結果を集約して、データ点を除去するために再訓練が必要なサブモデルを 1 つだけにする[Fredrikson 15]などである。

機械アンラーニングは、ユーザーがデータの一部を削除したいとの要望を出すと、その部分が最初から学習されていなかったかのようなモデル分布を生成するために、学習済みのモデルを再訓練する。

Bourtoule ら[Bourtoule21]は，学習アンラーニング問題を，サービス提供者 とユーザ集団 の二者間のゲームの一種として定式化している．Guo ら[Guo20]も、精度評価の観点から、同様の概念である検証削除(certified removal)を提案している。

しかし、研究者の中には、機械アンラーニングの定義について別の見解も与えている。Thudi ら[Thudi22]は、機械アンラーニングは厳密アンラーニング[Ullah21]と近似アンラーニング[Mahadevan21]に分けられるべきだと主張する。現在の機械アンラーニングの定義は、近似アンラーニングの出力を厳密アンラーニングの出力に限りなく近づけようとするものが主流である。しかし、彼らの批判は、異なるデータセットで学習しても同じモデルが得られる可能性もあり、この定義は正しくないと指摘する。また、この定義はアルゴリズムレベルでのみ適用されるという制限もある。

6.2 厳密なアンラーニング

厳密なアンラーニング[Ullah21]とは、予測タスクのように機械学習モデルを構築するためにユーザデータを直接利用する場合、ユーザデータが全くない場合の状態にシステムの状態を調整することが妥当な判断基準であることを意味する。
Ullahh ら[Ullah21]は、凸リスク最小化問題に対して、次の 3 つの性質を満たす全変動安定性(total variation stability)と呼ぶアルゴリズムであるを提案した。
- ストリームにおいて、各時点で、出力モデルは更新されたデータセットで学習した場合に得られるものと区別がつかないこと。
- 学習解除手法の実行時間が短いこと。
- 出力モデルの精度が高いこと。

厳密なアンラーニング手法は，例えばサポートベクター機械[Kashef21]，ナイーブベイズ[Jose21]，協調フィルタリング，リッジ回帰などに対して提案されている。ここでは，2 つの代表的な厳密なアンラーニングの手法を紹介する。

6.1.1. 最初の研究

Cao と Yang は[Cao15]で初めて機械アンラーニングの概念を導入した。彼らはモデル学習アルゴリズムを統計的問い合わせ（SQ）学習[Kearns98]に従った総和形式に変換する方法を提示している。このアンラーニング法は、学習データセットから少数の総和を更新するだけで実行される。少数の総和は、機械学習アルゴリズムとモデルの学習データの間のレイヤーに設定され、学習アルゴリズムは総和にのみ依存することで、更新の影響を抑えている。

著者らは、非適応型 SQ 学習（アルゴリズム開始前に全ての SQ が前もって決定される場合）と適応型 SQ 学習（後の SQ が前の SQ 結果に依存する場合）に基づいて、提案法を実装評価している。この場合、それらの総和形式は多くの機械学習モデルや全てのステージで実装できる汎用性も特徴である。

6.1.2. SISA トレーニングの考え方

Bourtoule ら[Bourtoule21]は、Sharded, Isolated, Sliced, and Aggregated training の略である SISA 学習法を提案している。この枠組みでは、学習手順においてデータポイントの影響を戦略的に制限することにより、学習解除プロセスを促進する。

このモデルでは、元のデータセット D を複数のサブデータセットに分割（Sharded）し、機械学習ネットワークも同数のサブネットワークにスライス(Sliced)し、それぞれのサブデータセットを対応するサブネットワークで学習し、最終的に学習結果を集約アルゴリズムで統

合(Aggregate)する。この方式では、データの一部削除が要求された場合、対応する特定のサブデータセットから、そのデータを削除し、局所的に再訓練を行うだけでよい。最後に、学習結果を再統合し、新しい学習結果を得る。こうして、不要なデータやモデルの学習処理を減らし、機械の再訓練にかかる時間や計算量を削減する。

原論文[Bourtoule21]では、単純な学習課題に対しては、SISA 学習法が、モデルの精度に影響を与えることなく、学習解除の要求を迅速に達成できることを示している。しかし、複雑な学習課題に対しては、モデルの精度への影響を軽減し、学習解除要求を迅速に完了させるためには、SISA 学習法を転移学習などの他の学習手法と組み合わせる必要があるとも課題付けている。

6.3 近似アンラーニング

近似アンラーニングとは，機械学習モデルやデータセットを調整することで、モデル再訓練の効果を近似的に向上させる方法である。Mahadevan ら[Mahadevan21]は、近似アンラーニングの手法は大きく 3 つに分離している。この節では、これを基に、幾つかくの既存手法を紹介する。

6.2.1 第一のカテゴリー

最初のカテゴリーは，残りのデータで機械学習モデルを再訓練し、フィッシャー情報行列の原理に基づいて最適なノイズを注入することで更新し，検証性を制御する[Martens20]。

差分プライバシー[Dwork14,Chaudhuri08]は、学習済みモデルのパラメータが個人情報を漏らさないことを保証することができる。Golatkar ら[Golatkar20a,Golatkar20b]は、差分プライバシーの手法に基づき、データセットの学習を選択的に解除し、機械学習モデルを更新する方法を提案した。彼らは、深層学習ネットワークの学習に使用した特定の学習データを削除するために、重みを「スクラビング」する方法を提案する。この方法では、ゼロからの再訓練や、最初に訓練した使用したデータへのアクセスは必要としない。その代わり、重みの任意のプローブ関数が、これらの特定のデータで訓練されていないネットワークの重みと同じ関数に近似するように、モデルの重みを変更する必要がある。

Golatkar ら[Goltkar21]は、それまでの研究を基に、機械アンラーニングに対して、新しい概念である混合プライバシー設定を導入している。このアイデアによれば、学習サンプルの「コア」部分集合は学習解除する必要がない。既存研究[Golatkar20a,Golatkar20b]と同様に、この新手法では、重みの部分集合をゼロに設定するだけで、非コアデータに含まれる全ての情報を最小限の性能低下で効果的に除去することができる。彼らは、この方法が大規模な画像分類タスクにおいて、精度および保証における非学習の大幅な改善をもたらすことを実証している。

6.2.2 第二のカテゴリー

次のカテゴリーは、機械アンラーニングの間に削除されたデータで機械学習モデルを更新し、ニュートン法[Koh 17]を実行して削除されたデータがモデルに与える影響を推定し、それを削除する。Guo ら[Guo20]は、ニュートン法のステップを 1 回行うことで近似的に再訓練を試み、この方法が問題解決に如何に効果的であるかを論じている。

Izzo ら[Izzo21] は，モデルの予測値に対する特定の学習点の影響度を評価するために，影響（Influence） 法[Koh 17, Giordano 19]を導入したアンラーニング法を提案している。提案手法は、計算コストが特徴次元の平方程度であり、学習データ数には依存しない。さらに、この方法は、線形回帰モデルにもロジスティック回帰モデルにも適用可能である。

6.2.3. 第３のカテゴリー

最後のカテゴリーは、機械学習モデルの学習中にデータと関連情報を保存し、データの削除要求があったときにモデルを更新するためにそれらを使用するものである[Wu20, Neel21]。

Graves ら[Graves21]は記憶喪失型(Amnesiac) というアンラーニング概念を提案し、モデル所有者が学習過程で機密性の高いデータやパラメータをバッチの形で保存しておくとする。データ削除の要求があった場合、モデル所有者は削除されたデータを含むバッチのパラメータ更新は行わない。このプロセスは、機密データを含む特定の機械学習ステップを選択的に取り消すと解釈することもできる。モデル学習は、初期モデルパラメータに対する一連のパラメータ更新と見なすことができる。このアプローチには、モデル所有者が機密データと関連するパラメータを保存するために大量の記憶領域を必要とするという潜在的な欠点がある。しかし、このスペースコストは厳密アンラーニング手法の計算量や処理時間よりは、はるかに少ないと著者らは主張している。

6.2.4. 評価基準

近似アンラーニングでは、データ削除のための効果的で高速なアルゴリズムの設計に加えて、近似アンラーニング手法の品質を適切に評価することが重要な課題である。そのため、多くの研究者がアルゴリズムの効果的な評価指標を提案している。Mahadevan ら[Mahadevan21]は、異なる学習法の性能を有効性、証明性、効率性の観点から測定するために、次の 3 つの評価指標を定義している。

-有効性：機械学習モデルの予測精度を測定する。

-検証性：更新されたモデルがどの程度遅延データを学習解除したかを測定する。
-効率：更新モデルと完全再訓練モデルを得るためのアルゴリズム実行時間の性能比を測定する。

Izzo ら[Izzo 21]は、近似データ削除法の有効性を評価するために、L2 距離と特徴注入テストの 2 つの指標を導入した。

- 更新モデルと完全再学習モデルの間の L2 距離は,近似アンラーニングの精度を測定するために用いられる比較的一般的な方法である。L2 距離の値が小さいほど、更新されたモデルの予測能力が完全再学習モデルの予測能力に近いことを示す。

- 特徴注入テストは、モデル（更新モデルと完全再訓練モデル）が学習を期待する残りのデータセットに、強い信号（追加特徴）として注入される。この特定の特徴を除去する前と後のモデルの学習パラメータの性能を観察することで、近似的な除去法の有効性を測定する。

7. データリネージとの結合

機械学習モデルのプライバシーを保護するプロセスにおいて、データフローの追跡は不可欠な要素である[Zhang17,Luo21,Thuago20]。

本節では、機械学習モデルのプライバシー保護におけるデータリネージ管理の役割について議論する[Zhang22a,Zhang22b]。

データリネージは、ソースシステムから様々な形態の永続化や変換を経て、最終的にアプリケーションや分析モデルによるデータ消費に至るまでのデータの動きを時系列で追跡する。データリネージ管理システムは、任意の時点で発生した機械学習モデルのあらゆるデータ変更を監視することができる[Li2022]。したがって、データリネージ管理システムとの組み合わせにより、機械学習モデルのセキュリティ保護を効果的に強化することが可能となる。

データ系統管理は、完全性攻撃と見なせるデータポイズニング攻撃など、特定のサイバー攻撃を防御するために適用できる。攻撃者は、学習データを改ざんすることで、モデルが正しい出力を予測することに影響を与える。攻撃者の目標は、自分の入力がモデルの学習データとして受け入れられるようにすることである。

Baracaldo ら[Baracaldo18]は、機械学習モデルの学習データセットに含まれるデータポイントのソース、変換、宛先に関するシステムの系譜をフィルタリングアルゴリズムの一部として使用することで、有毒データを識別することを提案し、これは原因攻撃の検出としても利用している。この手法により、オンラインかつ定期的に再訓練された機械学習システムは、潜在的に敵対的な環境におけるデータソースを識別することができる。最後に、この手法を応用して、Internet of Things 環境でも有毒なデータ注入を識別可能と報告している。

Google の Tensorflow チームは、機械学習データの系統管理のためのバージョン管理プラットフォーム、Machine Learning Metadata (MLMD)を開発し、公開している[Google22]。MLMD は、メタデータ、データ前処理、特徴選択、モデル学習、予測、評価、デプロイメントなど、機械学習ワークフロー全体の完全なデータ系統を追跡するためのライブラリと見なすことができる

MLMD は様々な機械学習パイプラインに実装することができ、モデルの学習中に生成された全てのデータを記録し、管理することができる。パラメータの更新やエラーのデバッグなど、モデルのデータ変換をすべて分析することができる。さらに、セキュリティの観点から、このメタデータプラットフォームは、機械アンラーニングとデータリネージを組み合わせた研究のためのアイデアも提供している。

機械アンラーニングの実装でも、データリネージは重要な役割を果たす。例えば、機械学習モデルの学習に使用した機密性の高い個人情報をユーザーが撤回したい場合や、悪意のあるデータ挿入攻撃を検知した場合など、開発者が機械アンラーニングを適用できる。学習用データの一部をデータセットから削除し、モデルを再学習させる必要があるが、データリネージ管理システムを使えば、これらの処理も漏れなく記録できる。

8. 今後の課題

機械アンラーニング分野の研究は,まだ 10 年足らずで、研究者はまだ多くの課題に直面している。例えば、ビッグデータ環境において機械アンラーニングが大量のデータ削除をいかに効率的に処理するか、機械学習プラットフォームがプライバシー侵害に遭遇した際にいかに迅速に対応できるか、データリネージ管理システムが機械学習のプライバシー保護の側面をどのように生かすことができるかなどが挙げられる。ここでは、そのうちの 4 つを考察する。

(I) 能動的・受動的なアンラーニング

今回取り上げた機械アンラーニングの手法は、いずれもデータ保有者の意思で能動的にアンラーニングを行うものである。しかし、機械学習モデルの CIA 特性に対しては、受動的なアンラーニングも有効な選択肢となる。攻撃者が機械学習モデルに対して CIA 攻撃を行った場合、データ保有者や機械学習プラットフォームがこの攻撃者の行動を発見するのが間に合わず、機械アンラーニング手法などのプライバシー保護対策を行う前に個人情報が漏洩してしまうことになる。この場合、受動的アンラーニング手法では、機械学習システムが攻撃された

時点でデータの削除が間に合うので、データ保有者の損失を最小限に抑えることができる。

（II）アンラーニングにおけるプライバシーのリスク

機械アンラーニングの本来の目的は、機械学習によるプライバシー漏えいを防ぐことである。しかし、近年、一部の研究者は、機械アンラーニングのプライバシー保護効果に疑問を投げかけている。Chen ら[Chen21]は、メンバーシップ推論攻撃[Shokri17, Sablayrolles19]の特定のシナリオにおいて、機械アンラーニング手法も攻撃されモデルのプライバシーが漏えいする可能性があると指摘している。彼らは、新しいメンバーシップ攻撃を設計し、機械学習モデルをゼロから再訓練するアプローチと SISA アプローチの 2 つの機械アンラーニング法に対して実験的な解析評価を行った。その結果、彼らの攻撃方法は、ゼロからの再訓練に大きな影響を与えること、一方、SISA のような分散型学習モデルに対しては、それほど影響を与えないことを報告している。

機械アンラーニングに対する攻撃と耐性評価は、まだ始まったばかりである。最新の攻撃論文[Marchant22]では、データ削減の計算コストを意図的に増加させる攻撃モデルが議論されている。

（III)データリネージを利用した解決策

機械学習のセキュリティとプライバシー保護のレベルでは、データリネージ管理システムは、すべてのデータとモデルの変更を追跡することができる。機械学習モデルの保護機構として機械アンラーニングアプローチが導入されたことで、データリネージマネジメントシステムとの併用が必須となった。機械アンラーニングは、同じ環境での学習に使用する機械学習モデルとは独立した別のモデルであると考えることができる。機械アンラーニングのデータ変更は、学習用機械学習モデルのセキュリティに直接影響するため、データリネージマネージメントシステムで、その履歴を記録しておく必要がある[Zhang22b]。

（IV）ユーザ側での検証

忘れられる権利の実行は、データ管理者に対しては、強制できるが、データ保有者自身が、積極的に規制する機能がない。つまり、データ保有者がどのようにプライバシーの侵害に気づくか、どのような場合に個人データの削除を要求するかということが課題であることは、本解説の第 2 章でも論じた。これに対する 1 つの技術解として、Eisenhofer ら[Eisenhofer21]は、データ保有者や利用者自身でも、検証可能な機械アンラーニングの枠組みを与えている。彼らの提案では、データの削除や更新、再訓練など一連の計算過程の証明に、暗号仮想通貨でも利用されている SNARK(Succinct Non-interactive ARgument of Knowledge) と呼ばれる高度な非対話型ゼロ知識証明技法を駆使する。この手法を使えば、利用者側には、必要最小限の事実が伝わるだけで、計算過程の詳細は秘密にできる。最後に、線形回帰、ロジスティック回帰、およびニューラル ネットワークでの機械アンラーニングに対する構築の実用性までも評価している。ただし、計算の正当性に関する証明の構成は、あくまでデータを管理する機械アンラーニングのシステム側であり、ユーザーであるデータ保持者は、受け身であることに注意する。

9. 最後に

機械学習におけるプライバシー保護アプローチとしては、準同型暗号を駆使した秘匿計算や、最近では連合学習でもプライバシーが主要課題の 1 つとして注目され、国内でも研究と実用化が盛んである。しかし、プライバシー保護技術に関する解説や調査でも、国内では未だ、機械アンラーニングへの言及までには至っていない現状にある。

プライバシーを保護する機械学習の研究は止まることを知らない。数あるアプローチの中でも、機械学習アルゴリズム自体と密接に関連する機械アンラーニングの研究が台頭してきている。現在の機械アンラーニングの研究は既に、差分プライバシー[Roth21]や連合学習[Wu22]など、様々な研究アプローチと融合して発展している。

本稿でも、機械学習におけるプライバシー保護手法を広くとらえて、機械アンラーニングもその一つの研究領域として解説した。また、学術論文では、アルゴリズムの計算時間として評価されている多くのコストは、実用現場では、再訓練などに要する経費に換算され、データ規模によっては、膨大な額になる場合があることにも注意する。

この解説を通じて、機械アンラーニングの理解と研究の活性化のきっかけとなれば幸いである。

謝辞: 本研究は、科学技術振興機構（JST）戦略的国際共同研究推進事業（SICORP）の一部助成を受けたものである。筆頭著者は JST SPRING、助成番号 JPMJSP2136 の支援を受けた。データリネージと機械アンラーニングに関しては、KDDI 総合研究の磯原氏と中村氏には、本研究の下地となった予備調査[Zhang22a, Zhang222b]で有益な議論を頂いた。北九州市立大学の姚智華博士には、草案編集に協力を頂いた。

◇参考文献 ◇

[Cambridge22] Cambridge Learner’s dictionary (Online)https://dictionary.cambridge.org/ja/dictionary/learner-english/ (2022.Jan. 03, Access)

[人事労務用語 22]人事労務用語辞典(online), https://kotobank.jp/word/アンラーニング-178709 (2022.Jan.03/ access)
[GDPR16] REGULATION (EU) 2016/679 OF THE EUROPEANPARLIAMENT AND OF THE COUNCIL (https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A32016R0679)(27 April 2016)
[MIT2022] Heikkiläarchive,M., Artists can now opt out of the next version of Stable Diffusion, MIT Technical Review,https://www.technologyreview.com/2022/12/16/1065247/artists-can-now-opt-out-of-the-next-version-of-stable-diffusion/ (December 16, 2022)
[Cao15] Cao, Y., Yang, J., Towards making systems forget with machine unlearning, 2015 IEEE Symposium on Security and Privacy, pp. 463-480. IEEE (2015)
[Zhou22] Zhou.J, Top Cyber Security Conferences Ranking http://jianying.space/conference-ranking.html (2022)
[Warnecke 23] Warnecke A., Pirch L., Wressnegger C. and Rieck K., Machine Unlearning of Features and Labels, arXiv:2108.11577 (2021) Accepted in NDSS 2023
[Mercuri 22] Mercuri S., Khraishi R., Okhrati R., Batra D., Hamill C., Ghasempour T., Nowlan A.,An Introduction to Machine Unlearning" https://arxiv.org/abs/2209.00939 (2022)
[Nguyen22] Nguyen T.T, Trung Huynh T.T.,Nguyen P.L., Liew A.W., Yin H., Nguyen Q.V.H , A Survey of Machine Unlearning https://arxiv.org/abs/2209.02299 (2022)
[16][披田野 16]披田野清良,村 隆夫,清本 晋作,花岡 悟一郎, ポイゾニングを利用したモデル再構築によるセンシティブ情報の復元に関する一考察, IPSJ CSS 2016 2A1-4 (2016)
[Zhang22a] Zhang H. Nakamura T. Takamasa I., SAKURAI.K, Data Lineage Management with Unlearning Method for Machine Learning Security and Privacy Issues. IEICE SCIS2022, 1D-41, (2022).
[Zhang2b] Zhang H. Nakamura T. Takamasa I., SAKURAI K., A review on Machine Unlearning, Submitted to a journal (2022)
[Baracaldo17] Baracaldo, N., Chen, B., Ludwig, H., Safavi, J.A.: Mitigating poisoning attacks on machine learning models: A data provenance based approach. Proc. of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 103-110 (2017)
[Liu22] Liu, Y., Fan, M., Chen, C., Liu, X., Ma, Z., Wang, L., Ma, J.: Backdoor defense with machine unlearning. INFOCOM 2022 - IEEE Conference on Computer Communications,pp.280-289 (2022)
[Bourtoule 2021] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C.A., Jia, H., Travers, A., Zhang, B., Lie, D., Papernot, N.: Machine unlearning., 2021 IEEE Symposium on Security and Privacy (SP), pp. 141-159 (2021).
[Al-Rubaie19] Al-Rubaie, M., Chang, J.M.: Privacy-preserving machine learning: Threats and solutions. IEEE Security & Privacy 17(2), 49-58 (2019)
[Shelter 21]Schelter, S.: Towards efficient machine unlearning via incremental view maintenance, Workshop on Challenges in Deploying and Monitoring ML Systems at the International Conference on Machine Learning (2021).
[Graves21] Graves, L., Nagisetty, V., Ganesh, V.: Amnesiac machine learning. Thirty-Fifth AAAI Conference on Artificial Intelligence (2021)
[Chen21] Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., Zhang, Y.: When machine unlearning jeopardizes privacy. Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, pp. 896-911 (2021)
[Gao 22] Gao, J., Garg, S., Mahmoody, M., Vasudevan, P.N.: Deletion inference, reconstruction, and compliance in machine (un) learning.Proc. Priv. Enhancing Technol. 2022(3): 415-436 (2022)
[Marchant 22] Marchant, N.G., Rubinstein, B.I., Alfeld, S.: Hard to forget: Poisoning attacks on certified machine unlearning.The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22) (2022)
[Baracaldo18] Baracaldo, N., Chen, B., Ludwig, H., Safavi, A., Zhang, R.Detecting poisoning attacks on machine learning in iot environments. 2018 IEEE International Congress on Internet of Things (ICIOT), pp. 57-64, (2018).
[Chundawta 2022] Chundawat, V.S., Tarun, A.K., Mandal, M., Kankanhalli, M.: Zero-shot machine unlearning. arXiv preprint arXiv:2201.05629 (2022)
[Toreini2020] Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G., Van Moorsel, A.: The relationship between trust in ai and trustworthy machine learning technologies., Proc. 2020 Conference on Fairness, Accountability, and Transparency, pp. 272-283 (2020)
[Surma20] Surma, J.: Hacking machine learning: towards the comprehensive tax- onomy of attacks against machine learning systems. In: Proceedings of the 2020 the 4th International Conference on Innovation in Artificial Intelligence, pp. 1-4 (2020)
[Tramer16] Tramer, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction apis. In: 25th {USENIX} Security Symposium ({USENIX} Security 16), pp. 601-618 (2016)
[Song17] Song, C., Ristenpart, T., Shmatikov, V.: Machine learning models that remember too much. In: Proceedings of the 2017 ACM SIGSAC Confer- ence on Computer and Communications Security, pp. 587-601 (2017)
[Abadi16]Abadi M., A. Chu A.,Goodfellow I.,McMahan H.B.,Mironov I.,Talwar K., and Zhang L., "Deep learning with differential privacy",2016 ACM SIGSAC Conference on Computer and Communications Security, pp.302-318, 2016.
[Shen16] Shen, S., Tople, S., Saxena, P.: Auror: Defending against poisoning attacks in collaborative deep learning systems. In: Proceedings of the 32nd Annual Conference on Computer Security Applications, pp. 508-519 (2016)
[Alsdurt20] Alsdurf, H., Belliveau, E., Bengio, Y., Deleu, T., et al.: COVI white paper. arXiv preprint arXiv:2005.08502 (2020)
[Ginart 2019] Ginart, A., Guan, M.Y., Valiant, G., Zou, J.: Making ai forget you: Data deletion in machine learning. arXiv preprint arXiv:1907.05012 (2019)
NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems, Article No.: 316Pages 3518-3531 (2019)
[Fredrikson 15] Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322-1333 (2015)
[Mahadevan21] Mahadevan, A., Mathioudakis, M.: Certifiable machine unlearning for linear models. arXiv preprint arXiv:2106.1509 (2021)

[Dwork2014] Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. 9(3-4), 211-407 (2014)
[Chaudhuri 2008] Chaudhuri, K., Monteleoni, C.: Privacy-preserving logistic regression. Advances in neural information processing systems 21 (2008)
[Guo20] Guo,C., Goldstein, T., Hannun, A., Van Der Maaten, L.: Certified data removal from machine learning models. Proc 37th International Conference on Machine Learning,No.359,pp 3832-3842 (2020)
[Thudi22] Thudi, A., Jia, H., Shumailov, I., Papernot, N.: On the necessity of auditable algorithmic definitions for machine unlearning. 31st USENIX Security Symposium,USENIX Association、 pp.4007--4022, (2022)
[Ullah21] Ullah, E., Mai, T., Rao, A., Rossi, R.A., Arora, R.: Machine unlearning via algorithmic stability. In: Conference on Learning Theory, pp. 4126-4142 (2021).
[Cao18] Cao, Y., Yu, A.F., Aday, A., Stahl, E., Merwine, J., Yang, J.: Efficient repair of polluted machine learning systems via causal unlearning. In: Proceedings of the 2018 on Asia Conference on Computer and Communications Security, pp. 735-747 (2018)
[Kashef 21] Kashef, R.: A boosted svm classifier trained by incremental learning and decremental unlearning approach. Expert Systems with Applications 167, 114154 (2021)
[Jose 21] Jose, S.T., Simeone, O.: A unified pacbayesian framework for machine unlearning via information risk minimization. In: 2021 IEEE 31st Inter- national Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6 (2021)
[Golatkar 20a] Golatkar, A., Achille, A., Soatto, S.: Eternal sunshine of the spotless net: Selective forgetting in deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9304-9312 (2020)
[Wu20] Wu, Y., Dobriban, E., Davidson, S.: Deltagrad: Rapid retraining of machine learning models. International Conference on Machine Learning, pp. 10355-10366 (2020).
[Goltkar21] Golatkar, A., Achille, A., Ravichandran, A., Polito, M., Soatto, S.: Mixed- privacy forgetting in deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 792-801 (2021)
[Izzo 21] Izzo, Z., Smart, M.A., Chaudhuri, K., Zou, J.: Approximate data deletion from machine learning models. In: International Conference on Artificial Intelligence and Statistics, pp. 2008-2016 (2021). PMLR
[Neel21] Neel, S., Roth, A., Sharifi-Malvajerdi, S.: Descent-to-delete: Gradient- based methods for machine unlearning. In: Algorithmic Learning Theory, pp. 931-962 (2021)
[Kearns98] Kearns, M.: Efficient noise-tolerant learning from statistical queries. Journal of the ACM (JACM) 45(6), 983-1006 (1998)
[Martens20] Martens, J.: New insights and perspectives on the natural gradient method. The Journal of Machine Learning Research (JMLR), 21(146):1-76, 2020.
[Dwork2014] Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. 9(3-4), 211-407 (2014)
[Chaudhuri 2008] Chaudhuri, K., Monteleoni, C.: Privacy-preserving logistic regression. Advances in neural information processing systems 21 (2008)
[Golatkar20b] Golatkar, A., Achille, A., Soatto, S.: Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In: European Conference on Computer Vision, pp. 383-398 (2020). Springer
[Koh 17] Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International Conference on Machine Learning, pp. 1885- 1894 (2017).
[Giordano19]Giordano, R., Stephenson, W., Liu, R., Jordan, M., Broderick, T.: A swiss army infinitesimal jackknife. The 22nd International Conference on Artificial Intelligence and Statistics, pp. 1139-1147 (2019).
[Zhang17] Zhang, Z., Sparks, E.R., Franklin, M.J.: Diagnosing machine learning pipelines with fine-grained lineage. In: Proceedings of the 26th International Symposium on High-Performance Parallel and Distributed Computing, pp. 143-153 (2017)
[Luo21] Luo, G., et al.: A roadmap for automating lineage tracing to aid automatically explaining machine learning predictions for clinical decision support. JMIR Medical Informatics 9(5), 27778 (2021)
[Thuago20]Thiago, R.M., Souza, R., Azevedo, L., Soares, E.F.D.S., Santos, R., Dos Santos, W., De Bayser, M., Cardoso, M.C., Moreno, M.F., Cerqueira, R.: Managing data lineage of o&g machine learning models: the sweet spot for shale use case. In: First EAGE Digitalization Conference and Exhibition, vol. 2020, pp. 1-5, European Association of Geoscientists & Engineers (2020)
[Li2022] Li, Y., Zheng, X., Chen, C., Liu, J.: Making recommender systems forget: Learning and unlearning for erasable recommendation. arXiv preprint arXiv:2203.11491 (2022)
[Google22] Google, *ML Metadata (MLMD)* is a library https://github.com/google/ml-metadata (2022. Accessed 2023.Jan)
[Shokri17] Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3-18 (2017).
[Sablayrolles19] Sablayrolles, A., Douze, M., Schmid, C., Ollivier, Y., J ´egou, H.: White- box vs black-box: Bayes optimal strategies for membership inference. In: International Conference on Machine Learning, pp. 5558-5567 (2019).
[Marchant 22] Marchant N.G,Benjamin I. P. Rubinstein B.I.P, Alfeld S., Hard to Forget: Poisoning Attacks on Certified Machine Unlearning:
Vol. 36 No. 7: AAAI-22 Technical Tracks 7 / AAAI Technical Track on Machine Learning II (2022)
[Eisenhofer21] Eisenhofer T.,Riepel D., Chandrasekaran V.,Ghosh E., Ohrimenko O., Papernot N., Verifiable and Provably Secure Machine Unlearning, arXiv:2210.09126 (2021)
[Roth21] Roth A., Gupta V., Jung C., Neel S., Sharifi S. Waites C.. Differential Privacy and Machine Unlearning.NeurIPS 2021 Workshop - Privacy in Machine Learning (PriML) 2021
[Wu22] Wu L., Guo S., Wang J., Hong Z., Zhang J., Ding Y., Federated Unlearning: Guarantee the Right of Clients to Forget, IEEE Network, Vol.36 Issue 5, pp.129-135, (2022)

著 者 紹 介

張 海波（非会員）

HAIBO ZHANG は、2015 年に中国・安徽大学でソフトウェア工学の学士号を取得、2018 年に米国・南カリフォルニア大学 Viterbi 工学部でサイバーセキュリティ工学の修士号を取得。現在は九州大学大学院システム情報科学府、博士後期課程。研究テーマは、IOT セキュリティ、デジタルサプライチェーンセキュリティ、敵対的生成ネットワークなど。

櫻井 幸一（会員）

1986 年九州大学・理学部・数学科卒業，1988 年九州大学工学研究科応用物理学専攻修士課程了．現在，九州大学大学院システム情報科学研究院情報学部門教授．2018 年より 国際電気通信基礎技術研究所（ATR）適応コミュニケーション研究所 客員研究員．2022 年より人工知能学会・安全とセキュリティ研究会主査．日本数学会，応用数理学会，電子情報通信学会，情報処理学会，IACR，ACM，IEEE 各会員．