机器学习的遗忘方式

班级-学号

摘 要 随着人工智能和机器学习的广泛应用，数据隐私问题愈加引人关注。机器学习遗忘是一种创新的隐私保护技术，旨在从模型中删除特定数据的影响，以满足用户“被遗忘权”的需求。本文综述了机器学习遗忘的基本原理与主要方法，包括模型无关方法、模型内在方法和数据驱动方法。本文还探讨了遗忘验证的常用方法，如特征注入测试和信息泄露检测，以确保遗忘的有效性。本研究为未来的机器学习遗忘技术开发提供了理论和实践的参考。

关键词 机器学习遗忘；隐私保护；精确遗忘；近似遗忘；模型验证

How to Unlearn: Approaches for Machine Forgetting

Zhifeng Han[1)]

[1)] (School of Cyber Engineering, Xidian University, Xi’an 710071, China)

Abstract With the wide application of artificial intelligence and machine learning, data privacy issues have attracted more and more attention. Machine learning forgetting is an innovative privacy protection technique that aims to remove the impact of specific data from the model to meet the needs of the user's "right to be forgotten". This paper reviews the basic principles and main methods of forgetting in machine learning, including model-independent methods, model-intrinsic methods and data-driven methods. This paper also explores common methods of forgetting verification, such as feature injection testing and information leakage detection, to ensure the effectiveness of forgetting. This study provides a theoretical and practical reference for the future development of machine learning forgetting technology.

Key words Machine learning forgetting; Privacy preserving; Precise forgetting; Approximate forgetting; Model verification

1 引言

在现代信息社会中，数据隐私和安全问题日益受到关注。随着人工智能（AI）和机器学习（ML）技术的广泛应用，系统中积累的海量数据成为一种不可忽视的资源，但也带来了数据滥用和隐私泄露的风险。近年来，随着《通用数据保护条例》（GDPR）等法律法规的出台，数据隐私保护成为技术开发的强制性要求，这推动了“遗忘权”的发展。机器的遗忘，即“how to unlearn”，指的是机器能够主动或被动地删除已学习的信息，使其不再受特定数据影响。这一概念不仅在隐私保护中具有重要意义，也可用于数据清理、模型改进等多个领域。

1.1 机器遗忘的背景与必要性

在传统机器学习和深度学习模型中，一旦模型从数据中学得了某些特征或规律，这些知识便成为模型不可分割的一部分。这使得在模型训练完毕后移除某些数据的影响变得极为困难。尤其是在使用

敏感或隐私数据的应用场景中，遗忘这些数据的能力变得至关重要。例如，金融数据、医疗记录和社交媒体等领域中用户信息的保护要求，要求系统能够在用户撤回数据或进行隐私删除后保证相关数据对模型不再产生影响。

1.2 机器遗忘的主要挑战

机器学习遗忘技术在实现过程中面临诸多挑战。首先，训练过程中的随机性增加了删除特定数据影响的复杂性，使得追溯和移除指定数据的“记忆”变得困难。此外，由于模型训练的增量特性，删除某一数据样本可能会间接影响后续数据的学习表现，从而要求算法能够彻底消除这种扩散效应。进一步，灾难性遗忘问题使得模型在逐步删除大量数据后性能显著下降，甚至导致整体失效，这对深度学习模型尤为明显。同时，为确保数据彻底删除，验证过程需要复杂的隐私检测，但在大规模模型中往往伴随较高的计算开销。除此之外，完全重训虽能保证遗忘效果，但带来的计算和存储成本难以满足实际需求，尤其在需要频繁遗忘的场景中更为不便。遗忘方法的通用性不足、难以适应不同模型架构，且在处理连续数据删除请求时易受到恶意攻击，这些问题均制约了机器学习遗忘技术的广泛应用和发展。

例如，模型修剪通过移除冗余参数来减少特定数据的影响，适用于大型网络，但可能影响模型性能。数据反向传播清除则通过调整模型参数来减少数据影响，但其计算成本较高，适用性受到限制。知识蒸馏方法则能在一定程度上有效地删除数据影响，同时保留模型性能，但实现过程较为复杂，适用于特定场景。

1.3 本文的研究内容

本文主要研究了机器学习遗忘的基本原理与实现方法，详细探讨了如何在模型中删除特定数据的影响，以满足数据隐私保护的需求。本文首先介绍了机器学习遗忘的定义和目的，随后对不同的遗忘方法进行了分类，包括模型无关方法、模型内在方法和数据驱动方法，分析了它们的适用场景和实现机制。

2 相关工作

机器学习遗忘的目标是使模型在数据删除请求后不再包含被删除数据的影响，而实现这一目标的方法多种多样。现有的研究工作主要集中在以下几个方面：

2.1 差分隐私与数据删除

最早的隐私保护技术如差分隐私（Differential Privacy）被提出用于减少单个数据对模型的影响，从而保证个体数据不易被推断。Dwork 等人提出的差分隐私方法[1]通过在模型的参数或输出中加入噪声，使得单条数据的删除或添加不会显著改变模型的行为。然而，差分隐私主要关注的是防止数据的“记忆”产生过多影响，而非彻底删除数据对模型的影响。在一些应用场景中，这种噪声干扰方法可能导致模型精度下降，并且在处理大规模数据删除时效率不高。

数据删除技术通常仅限于删除数据库中的数据条目，但并未解决模型内化数据后难以删除的情况。为此，机器学习遗忘应运而生，作为一种超越数据删除的进一步技术，旨在从根本上移除数据的影响。

2.2 机器遗忘的框架与方法

机器学习遗忘不同于简单的数据删除，其目标是从模型的学习参数中删除指定数据的影响，而不需重训整个模型。为此，近年来的研究提出了不同的遗忘框架与方法。Guo 等人[2]提出了对数据删除效果进行认证的方法，通过影响函数估计数据对模型参数的影响并逐步消除这种影响。这种方法在理论上对线性模型和凸损失函数具有较强的保证，但在非线性模型（如深度学习）中则面临挑战。Golatkar 等人[3]进一步在深度神经网络中应用该方法，通过扰动模型权重以掩盖删除数据的影响，提出了适用于随机梯度下降（SGD）算法的上界计算方法。Gupta 等人[4]通过差分隐私框架设计了一种流式删除机制，支持连续的删除请求序列，同时确保数据删除后的隐私保护。该方法适用于非自适应的数据流删除序列，但无法保证自适应数据删除场景下的高精度。Golatkar 等人提出了一种基于噪声扰动的方法，通过向模型权重增加随机噪声来掩盖删除数据的影响[5]。该方法适用于深度神经网络，在使用梯度下降优化时尤为有效。另一种方法是通过对深度模型的隐藏层特征进行分离，将与待删除数据相关的特征逐步移除，从而实现数据遗忘[6]。一些研究表明，通过将特定数据的特征从模型的高层抽象表示中分离出来，可以有效实现数据的删除。例如，Guo 等人[7]提出了一种基于解耦表示的

方法，将特征与模型输出之间的关联进行量化，并在删除请求下逐步解除这些关联。统计查询学习方法通过统计数据样本特征而非原始数据进行模型训练，因此可以在移除样本后重新计算统计量以达到删除效果[8]。这种方法适用于大型数据集，但在深度学习中的应用受到限制。Koh 和 Liang 提出了影响函数法[9]，用于估计单个数据点对模型训练的影响。该方法通过影响函数直接计算待删除数据对模型参数的贡献，并逐步消除这种影响，尤其适用于线性模型或浅层模型。在深度学习中，影响函数的计算成本较高，因此在大规模模型中应用受到限制。

3 机器遗忘的基本原理

机器学习遗忘的核心在于通过算法使模型“忘记”特定的数据，保证删除数据的影响被移除。机器学习遗忘通常可以分为精确遗忘和近似遗忘两大类，这两种方式分别具有不同的实现机制和适用场景。

3.1 精确遗忘

精确遗忘（Exact Unlearning）的目标是通过删除特定数据使模型表现与从未见过该数据的情况相同。这一方式通常通过重新训练模型来实现，即模型在去除指定数据后的子集上重新训练，从而获得不包含该数据影响的模型。

给定训练数据集D和待删除的子集D_f，模型训练可以表示为学习算法A对数据集D的映射：

$$w = A(D) \tag{1}$$

其中，w为基于数据集D训练的模型参数。如果我们希望从模型中删除D_f的影响，精确遗忘的目标是使模型在删除后的数据集$D_r = D\backslash D_f$上中心训练，即：

$$w_r = A(D_r) \tag{2}$$

在精确遗忘的理想情况下，模型参数w_r应该等价于删除D_f后的模型参数$U\left(D, D_f, A(D)\right)$，其中$U$表示遗忘算法。因此，精确遗忘要求满足：

$$P_r\left(A\left(D\backslash \mathrm{D}_f\right)\right) = P_r\left(U\left(D, D_f, A(D)\right)\right) \tag{3}$$

这一过程保证模型不会在任何程度上“记住”删除的数据。尽管精确遗忘能够保证数据的彻底删除，但其高昂的计算成本使其在实际应用中难以推广。

3.2 近似遗忘

近似遗忘（Approximate Unlearning）通过对模型的参数或输出进行调整，在不完全重新训练的情况下实现近似遗忘效果。近似遗忘在效率上优于精确遗忘，适合应用在计算资源有限的场景中。近似遗忘可定义为以下约束条件：

给定$\epsilon > 0$，若遗忘算法U能够满足以下条件，则称其为ϵ-近似遗忘：

$$e^{-\epsilon} \le \frac{P_r\left(U\left(D, z, A(D)\right) \in T\right)}{P_r(A(D\backslash \mathrm{z}) \in T)} \le e^{\epsilon} \tag{4}$$

其中，z是被删除的单个样本，T是任意的事件集合。该条件通过给定ϵ范围，确保遗忘后的模型与未包含z的重新训练模型在分布上接近。此条件在实际应用中常用语评估遗忘效果的准确性与效率。

3.3 遗忘请求类型

在实际应用中，遗忘请求可能以不同形式出现，如移除单个样本（Item Removal）、特征移除（Feature Removal）、类别移除（Class Removal）等。为满足这些不同的需求，遗忘算法通常针对特定的数据类型进行设计。例如，特征移除通常涉及去除特定维度的数据影响，类别移除则需要在多分类任务中删除某一类别的样本并保持模型的整体性能。

3.4 遗忘机制的设计要求

为了实现有效的机器学习遗忘，遗忘机制通常需要满足以下设计要求。首先是完整性，即遗忘后的模型应与重新训练模型在输出上保持一致。其次是时效性，即近似遗忘方法应能够在合理时间内实现数据遗忘。最后是准确性，即保持模型精度的前提下尽可能实现有效的遗忘。

4 机器学习遗忘的方式

机器学习遗忘的方法可以根据其适用性和实现方式划分为三大类：模型无关方法、模型内在方法和数据驱动方法。这些方法各自适用于不同的应用场景，在资源消耗的遗忘精度上也有所不同，具体的各种机器学习遗忘方式的优缺点如表 1 所示。

4.1 模型无关方法

模型无关方法并不依赖于特定的机器学习模型架构，因此适用于多种类型的模型。

4.1.1 差分隐私（Differential Privacy）

差分隐私的核心思想是通过向模型的参数或

表 1 不同机器遗忘方法的优缺点对比

方法类别	主要方法	优点	缺点
模型无关方法	差分隐私、认证删除机制	通用性强，适用多种模型	噪声可能导致模型精度下降
模型内在方法	权重扰动、特征分离	适合深度学习，遗忘效果显著	在高精度任务中会降低性能
数据驱动方法	统计查询学习	计算高效，适合大规模数据集	在深度模型中应用有限

输出添加噪声，从而限制每个数据样本对模型的影响，减少模型“记住”单个数据的可能性。其主要目标是确保在数据样本被添加或删除时，模型的行为不会显著变化。

差分隐私的定义公式为：

$$P_r(A(D \cup \{z\}) \in T) \leq e^{\epsilon} P_r(A(D) \in T) \quad (5)$$

其中ϵ是隐私预算，控制数据对模型影响的限制范围。当ϵ越小，隐私保护越强，数据样本对模型的影响越小。差分隐私适用于需要加强隐私保护的场景，但添加噪声可能会降低模型的预测精度。

4.1.2 认证删除机制（Certified Removal Mechanisms）

认证删除机制通过计算影响函数来预估单个数据点对模型参数的影响，并逐步抵消这种影响，从而实现数据的删除。对于线性模型，认证删除机制通过影响函数的计算能够在理论上保证删除效果，而在深度神经网络中则可能需要结合噪声扰动等方法。影响函数的计算公式为：

$$I(z, w) = -H_w^{-1} \nabla L(z, w) \quad (6)$$

其中H_w是模型参数的海森矩阵，表示模型对数据的敏感度，$\nabla L(z, w)$是损失函数对数据样本 z 的梯度，$I(z, w)$表示数据样本z对模型参数的影响。

4.2 模型内在方法

模型内在方法通过直接调整模型结构或参数，以删除特定数据对模型的影响。这种方法在深度学习模型中尤为有效。

4.2.1 权重扰动（Weight Perturbation）

权重扰动方法通过对模型参数引入噪声来掩盖删除数据的影响，使得模型在执行遗忘后对该数据无记忆。其基本思想是将数据样本z的影响通过随机噪声平滑化，从而实现数据删除。其公式为：

$$w' = w + \eta \nabla L\left(D_f, w\right) \quad (7)$$

其中w是原始模型参数，w'是引入噪声后的新模型参数，η是控制噪声大小的步长，L是损失函数。

权重扰动方法子啊删除数据影响时不会显著改变模型的整体结构，但在高精度要求的任务中可能会引入误差，导致性能下降。

4.2.2 特征分离（Feature Separation）

特征分离通过对模型的隐藏层进行操作，将与待删除数据相关的特征从模型中隔离或遮蔽，以减少其对模型预测的影响。该方法通常用于深度神经网络中的卷积层或隐藏层。

假设模型输出为$f(x; w)$，而特征$h(x)$对应待删除数据，则特征分离的实现公式可以表示为：

$$f(x; w) = g\left(h(x), h_{D_f}(x); w\right) \quad (8)$$

其中$h_{D_f}(x)$是与待删除数据D_f相关的特征表示，g是用于特征组合的函数。特征分离在图像、文本等高维数据上较为有效，通过减少相关特征的影响实现数据的遗忘效果。

4.2.3 数据驱动方法

数据驱动方法利用数据本身的特性实现遗忘，尤其适用于需要高效删除大量数据的场景。

4.2.1.1 统计查询学习（Statistical Query Learning）

统计查询学习通过计算样本的统计特征来训练模型，而非直接依赖原始数据。删除数据后重新计算这些统计特征，可以实现近似的遗忘效果。

假设原始模型参数w是基于数据统计量 $S(D)$ 训练的，那么删除数据后，我们重新计算统计量$S\left(D \setminus D_f\right)$，并更新模型参数：

$$w' = A\left(S(D \setminus Df)\right) \quad (9)$$

统计查询学习适合在大规模数据集上应用，但在深度学习模型中应用受限，因为深度模型依赖于直接的样本表示。

5 结论

随着数据隐私保护需求的提升，机器学习遗忘已成为保障用户数据隐私的关键技术。本文通过系统地回顾机器学习遗忘的基本原理与实现方式，总

结了三大类主要遗忘方法：模型无关方法、模型内在方法和数据驱动方法，并分析了每种方法的优缺点。模型无关方法具有广泛的适用性，但在实际应用中可能引入精度损失；模型内在方法更适合深度学习模型，能够在不完全重训的情况下实现高效遗忘；而数据驱动方法则在大规模数据删除的场景下表现优越。

参 考 文 献

[1] Dwork C. Differential privacy: A survey of results. In: Proceedings of the 5th International Conference on Theory and Applications of Models of Computation. Xi'an, China, 2008: 1-19.

[2] Guo C, Goldstein T, Hannun A, van der Maaten L. Certified data removal from machine learning models. In: Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria, 2020: 3832-3842.

[3] 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. Seattle, USA, 2020: 9304-9312.

[4] Gupta A, Kumar A, Venkatasubramanian S. A general framework for deleting data from models trained with SGD. arXiv preprint arXiv:2005.04118, 2020.

[5] 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. Seattle, USA, 2020: 9304-9312.

[6] Golatkar A, Achille A, Soatto S. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In: Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 383-398.

[7] Guo C, Goldstein T, Hannun A, van der Maaten L. Certified data removal from machine learning models. In: Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria, 2020: 3832-3842.

[8] Cao Y, Yang J. Towards making systems forget with machine unlearning. In: Proceedings of the 2015 IEEE Symposium on Security and Privacy. San Jose, USA, 2015: 463-480.

[9] Koh PW, Liang P. Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 1885-1894.