电信科学
Telecommunications Science
ISSN 1000-0801,CN 11-2103/TN

《电信科学》网络首发论文

题目：面向机器学习的隐私保护关键技术研究综述
作者：刘姿杉，程强，吕博
收稿日期：2020-05-11
网络首发日期：2020-08-17
引用格式：刘姿杉，程强，吕博．面向机器学习的隐私保护关键技术研究综述．电信科学. https://kns.cnki.net/kcms/detail/11.2103.TN.20200814.1745.024.html

面向机器学习的隐私保护关键技术研究综述[1]

刘姿杉，程强，吕博

(中国信息通信研究院，北京，100191)

摘　要：随着信息通信技术的发展，机器学习已经成为多个研究领域与垂直行业必不可少的技术工具。然而，机器学习所需数据中往往包含了大量的个人信息，使其隐私保护面临风险与挑战，受到了越来越多的关注。对现有机器学习下隐私保护法规政策与标准化现状进行梳理，对适用于机器学习的隐私保护技术进行详细介绍与分析。隐私保护算法通常会对数据质量、通信开支与模型表现等造成影响，因此对于隐私保护算法的评估应当进行多维度的综合评估。总结了适用于机器学习应用的隐私保护性能评估指标，并指出隐私保护需要考虑对数据质量、通信开支以及模型准确率等之间的影响。

关键词：机器学习；隐私保护；评估指标

中图分类号：TP181

文献标识码：A

A survey on the privacy preserving machine learning

LIU Zishan , CHENG Qiang　LÜ Bo

China Academy of Information and Communications Technology, Beijing, 100191

Abstract: With the development of information and communication technology, large-scale data collection has vastly promoted the application of machine learning in various fields such as computer vision, nature language processing, intelligent robots, automatic driving, and has become indispensable in multiple research fields and vertical industries. However, the data involved in machine learning often contains a lot of personal private information, which makes privacy protection face new risks and challenges, and has attracted more and more attention. The current progress of the related laws, regulations and standards to the personal privacy preserving and data safety in machine learning were summarized. The existing work on privacy preserving machine learning was then presented in detail. Privacy preserving algorithms usually have influence on the data quality, model performance and communication cost. Thus, the performance of the privacy preservation algorithms should be comprehensively evaluated in multiple dimensions. The performance evaluation metrics for the privacy preserving algorithms for machine learning were presented, given with the conclusion that the privacy preservation on machine learning needs to balance the data quality, model convergence rate and communication cost.

Key words: machine learning, privacy preserving, performance metrics

1　引言

收稿日期：2020-05-11；
基金项目:国家重点研发计划资助项目　（No.2018YFB1801200）
Found Item:The National Key Research and Development Program of China（No.2018YFB1801200）

近年来，机器学习已经成为图像处理、语音识别、自然语言识别、自动推理等技术的核心工具，并在智能客服、用户画像、智能检测、自动驾驶等领域取得突破性应用。机器学习算法的应用离不开可用的数据，大规模的数据收集能够提高机器学习应用的性能，但与此同时，数据集中包含的许多个人隐私信息会随着数据集的共享与应用而面临泄露的风险。另一方面，机器学习过程本身面临隐私问题，例如攻击者通过与模型交互从而对其中训练数据的敏感属性进行逆向推理。因此实现隐私保护的机器学习，已经成为机器学习发展的基础，并开始受到研究界与工业界越来越多的关注。

尽管隐私的概念已经成为目前的热门话题，但至今尚未有普遍接受的标准定义[1]。隐私定义的困难在于隐私所涉及的领域包括信息、实体、通信、领土等多个分类。在信息范围内，参考文献[2]将隐私定义为“个人、团体或机构规定有关自身的信息在何时、何地以及以何种程度可以传达给其他人的权利”，换言之，就是控制个人信息使用的权利。在参考文献[3]中，作者明确定义了个人隐私是公民个人生活中不愿被他人（一定范围以外的人）公开或知悉的秘密。

研究者们就如何在机器学习的应用过程中保障隐私这一问题，已经展开了许多研究，也取得了一系列进展。为了防止数据发布、披露、挖掘以及面向机器学习应用等过程中的隐私泄漏，目前已经出现了很多隐私保护技术和相关的研究综述[4-8]。其中，参考文献[4-7]主要对机器学习在不同阶段面临的安全与隐私威胁，包括训练阶段的数据泄露、预测阶段的模型反演攻击、黑盒/白盒攻击、成员推理攻击、模型窃取攻击等以及对应的防御技术包括同态加密、差分隐私等进行了总结。参考文献[8]中则着重介绍了在机器学习和深度学习中应用差分隐私进行输入扰动、梯度扰动或目标扰动的方法，并增加了对联邦学习的介绍。

面向机器学习的隐私保护研究可以分为共享原数据下的隐私保护技术以及保留原数据进行本地训练的隐私保护技术两大类。一般来说，对原数据的隐私保护处理往往会降低数据的可用性；而在保留原数据进行本地训练的技术例如同态加密、联邦学习，尚存在应用场景受限、信息交互量大与训练延迟较大等限制。建立机器学习下的隐私保护性能评估体系，对不同机器学习应用场景下衡量与选择适用的隐私保护技术尤为重要。本文首先介绍机器学习的隐私保护技术，并从隐私质量、数据可用性、算法复杂度与模型性能等多个维度建立机器学习隐私保护技术评估体系，着重指出隐私保护与数据质量、资源消耗以及模型可解释性之间的互相影响。

2 法律法规与标准化现状

目前国内外在通用数据和个人隐私保护方面已有相关的法律法规与标准实行。对于面向人工智能和机器学习的数据安全和隐私保护也开始日益重视，并开始了相关的立法和标准化工作，但完善的法律法规和标准体系还尚未形成。

2.1 法律法规现状

数据安全与个人隐私保护目前已经受到各国越来越多的关注。2018 年欧盟正式强制执行的《通用数据保护条例》是目前关于个人信息保护方面最重要和全面的法规，指明个人拥有管理自己个人数据的权利，

并提出通过匿名化的方式来保护个人敏感数据。类似的个人数据保护法律法规还有美国的《2018 年加州消费者隐私法案》、澳洲的《隐私保护原则》、巴西的《个人数据保护法》、日本的《个人信息保护法》等。我国也高度重视数据安全与隐私保护，但在立法方面目前较为分散。针对数据应用过程中的个人信息保护问题，2012 年第十一届全国人民代表大会常务委员会通过了《全国人大常委会关于加强网络信息保护的决定》，该决定要求网络服务提供者和其他企事业单位应当采取一定的措施对设计公民个人身份与隐私的电子信息进行识别和保护，防止信息泄露损毁和丢失等。2016 年我国发布了《中华人民共和国网络安全法》，要求网络运营者采取数据分类、重要数据备份和加密等措施，防止网络数据被窃取或者篡改，加强对公民个人信息的保护，防止公民个人信息被非法获取、泄露或者非法使用等。除此之外，对数据隐私保护的相关要求在《民法总则》、《网络安全法》、《个人信息保护法》、《消费者权益保护法》等法律中也均有涉及。

近些年，世界多个国家均把人工智能发展提高到国家战略层面，并开始在数据安全与隐私保护方面进行相关政策建议与立法工作。2019 年 4 月，欧盟委员会发布了《可信赖人工智能伦理指南》，指出人工智能系统必须确保隐私和数据保护，并遵从《通用数据保护条例》的关键原则。美国 2019 年版《国家人工智能研发与发展战略计划》，要求审查各自联邦数据和模型，注重保护数据安全、隐私和机密性。我国高度重视人工智能发展与相关数据集的建设与风险防范，2017 年 7 月，国务院印发《新一代人工智能发展规划》，其中专门提出人工智能发展要强化数据安全与隐私保护。2019 年 6 月，国家新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》，将“尊重隐私”作为八项原则之一，要求人工智能发展应尊重和保护个人隐私。但总体来看，我国现今尚未形成完善的人工智能与机器学习隐私保护相关的法律法规体系，为加快人工智能与机器学习应用发展，需要建设统一完备的数据安全与隐私保护的法律政策体系。

2.2 标准化现状

国际标准化组织近年来积极开展面向人工智能与机器学习的数据隐私保护相关标准。国际上，IEEE 个人数据隐私（PDP）工作组于 2016 年开展 P7002 数据隐私处理项目，IEEE P3652.1 联邦学习基础框架与应用工作组已开展联邦学习相关的标准化工作。ISO 和 IEC 信息技术联合委员会（ISO/IEC JTC1）下属的安全技术分委员会 ISO/IEC JTC1 SC27 下设专门用于隐私保护技术研究的工作组（WG5），主要对通用数据法人隐私保护相关技术进行研究，目前已经发布的相关标准包括 ISO/IEC 29100:2011《信息技术　安全技术　隐私保护框架》、ISO/IEC 29101:2013《信息技术　安全技术　隐私保护体系结构框架》、ISO/IEC 29190:2015《信息技术 安全技术 隐私保护能力评估模型》、ISO/IEC 29191:2012《信息技术 安全技术　部分匿名、部分不可链接鉴别要求》等。

我国标准化组织也开始重视人工智能数据安全标准。全国信息安全标准化技术委员会（SAC/TC260）在人工智能应用、大数据、个人信息保护等领域开展了数据安全相关标准工作，包括《信息安全技术 大数据服务安全能力要求》，对与机器学习相关的大数据安全能力提出要求；《信息安全技术 个人信息安

全规范》，明确了个人信息的收集、保存、使用、共享的合规要求，为机器学习中的数据安全与隐私保护提供重要参考。中国通信标准化协会（CCSA）对人工智能应用中数据安全管理与个人信息保护提出要求，包括《人工智能终端产品 个人信息保护要求和评估方法》、《人工智能服务平台数据安全要求》等，并在数据安全保护领域专门成立特设组，来对数据分级分类、数据安全合规性要求等重要标准进行研究制定。中国电子技术标准化研究院牵头开展人工智能安全标准框架研究，将人工智能数据安全列为重要研究内容。

3 面向机器学习的隐私保护技术

机器学习的隐私保护不仅包括对原数据集的个人信息保护，还包括对于模型所携带信息的保护，能够抵御模型窃取攻击、模型反演攻击等。基于此，机器学习的隐私保护分为两大类，一种是对于原始数据的隐私保护；另一种是对于不分享原始数据的情况下对训练数据与模型的保护。

3.1 原数据隐私保护技术

对于原始数据的隐私保护一般是针对集中学习过程，数据拥有者的数据被收集到执行机器学习任务的数据应用方。数据拥有者的数据一旦被收集，就很难再拥有数据控制权。因此在数据进行分享发布前，需要对原数据进行隐私保护处理，一般有基于数据的限制发布以及基于数据失真两种原理。

（1） 数据限制发布技术

基于数据限制发布的技术通过将数据集里与个人敏感信息或攻击者背景知识相关的属性定义为准标识符，通过对数据集中准标识符值进行泛化、抑制和隐匿处理，从而降低发布的数据精度来保护个人信息与属性。目前已有的研究中，最著名的方案包括 *k*-anonymity[9]以及随后衍生出的 *l*-diversity[10]和 *t*-closeness[11]等方法。其中 *k*-anonymity 方法使得每条数据记录至少与数据表中其他$k-1$条记录具有完全相同的准标识符属性值，从而使得数据应用者或攻击方无法通过准标识符直接连接到独立的个体。*l*-diversity 方法通过泛化操作，确保匿名数据中每个等价类至少包含 *l* 个不同的敏感属性值，通过敏感属性的多样化保证单一个体的敏感值不会暴露。*t*-closeness 方法要求所有等价类中敏感属性值的分布与该属性的全局分布保持一致。用$P=(p_1,p_2,...,p_m)$表示原数据中属性值的分布，$Q=(q_1,q_2,...,q_m)$表示一个等价数据集中属性值的分布，如果$Dist(P,Q)\le t$，则Q满足 *t*-closeness。

以上基于数据限制发布的方法中通过设置不同的参数，例如 *k*-anonymity 技术中的 *k* 值，可以平衡隐私保护和数据价值。当数据提供方在组织内部利用数据进行机器学习应用时，可以设置较小的参数值。而当需要将数据提供给外部服务方时，可以选择较大的参数值。但上述的隐私技术都从泛化数据集群的角度去保护用户的身份和敏感数据，但无法评估单条数据记录在数据集中对个人隐私泄露的影响。同时以上隐私保护方法无法提供可量化的隐私保护水平，不同数据集分布下隐私保护水平可能存在较大差异，从而削弱了其面向机器学习应用的可靠性。

（2） 数据失真保护技术

基于数据失真的技术通过对原始数据进行随机化或加噪扰动，使得处理后的数据失去重构性，从而实

现在进行数据挖掘、数据分析等操作中的数据隐私保护。2006 年, Dwork 等[12]提出了差分隐私的概念，是目前数据失真保护技术中的代表性方法。为了实现差分隐私，通过在数据集中添加噪声，例如 Laplace 机制和指数机制，可以使得相差一个数据的两个数据集查询结果满足概率不可分。一种隐私保护算法或策略 $M(\cdot)$ 满足 ε-差分隐私的条件是，对于任意两个至多相差一条数据的数据集合 T 、T' 满足：$\Pr[M(T) \in S] \leq e^{\varepsilon}\Pr[M(T') \in S] + \delta$，即攻击者能够推测出的敏感泄露被限定在 ε 之内。ε用于控制算法的隐私保护程度，取值越小，隐私保护性能越好。

理论上，通过在数据中添加噪声扰动，总能实现差分隐私。面向机器学习的应用，不仅可以对原数据集进行输入扰动实现差分隐私，还可以对机器学习模型的目标函数、梯度、输出的模型参数以及机器学习算法的真实输出结果引入随机噪声，保证整个机器学习过程满足差分隐私。由于差分隐私满足后处理、隐私叠加性和群隐私等优点[13]，因此在机器学习过程需要对包含相同个人的多个数据源进行融合时，只要每个数据集能够满足差分隐私，在进行数据融合或多次访问后的数据集仍可满足一定的差分隐私特性。

目前业界已经出现了利用拉普拉斯噪声对支持向量机（support vector machine, SVM）的模型权重进行扰动的方法，来使得算法满足ε-差分隐私特性。但如果模型权重中噪音过多，会降低分类的准确度；通过在代价函数中添加拉普拉斯噪声，可以保证分类结果较高的准确度，但缺陷是必须要特定的代价函数。在进行深度学习时，也可以在隐私成本可控的情况下完成深层神经网络的训练，例如满足差分隐私的生成对抗网络 GANobfuscator，通过在学习过程中为梯度添加噪声来实现生成对抗网络（generative adversarial network, GAN）下的差分隐私[14]。Phan 等[15]针对自编码器和卷积深度置信网络提出将目标函数近似为多项式形式，进而对其进行噪声扰动，从而使训练过程满足差分隐私。但由于深度神经网络的最优参数求解是非凸问题，模型训练收敛慢，通过在目标函数添加扰动来实现模型差分隐私，或在每次参数更新时都实现差分隐私的操作会使得整个训练过程的隐私保护成本巨大，难以平衡数据隐私、模型收敛度及模型可用性。近年来，基于宽松差分隐私定义的隐私保护方法被应用于深度学习训练过程[16]，来解决差分隐私的局限性。

3.2 不分享原数据下的隐私保护技术

对于不分享原数据的情况，机器学习的训练过程一般采用同态加密、联邦学习等方式，通过对密文做运算，或是数据拥有者直接利用数据在本地进行模型训练保护原始数据。但前者增加了计算的复杂度，后者获得中心模型的过程仍面临诸多问题与挑战。

（1） 同态加密

同态加密是一种不需要访问数据本身就可以对数据做运算与分析的密码学技术，它支持对密文直接进行运算，将得到的结果解密后与直接在明文上的运算一致[17]。同态加密常与联邦学习相结合，通过进行分布式模型训练的中间结果进行加密，从而保证模型训练过程的隐私与安全性。然而传统的同态加密算法只支持加法与乘法的多项式运算，机器学习中对于幂运算等更为复杂的数据处理支持有限。同时，同态加密技术运算复杂度较大，容易导致训练过程缓慢。随着加密后的数据空间变大，为机器学习的计算和通信都

带来了巨大开销。

（2） 联邦学习

作为一种分布式机器学习框架，它允许用户利用本地数据集进行模型的训练，在训练的过程中，数据本身不会离开用户本地，并通过加密机制下的参数交换来保证数据隐私，从而保证了原始数据的隐私保护[18]。然而，联邦学习过程仍然面临其他隐私问题，参考文献[19]对此进行了较为全面的总结分析。联邦学习可以利用差分隐私、同态加密等技术进行分布式训练过程的隐私保护增强[20]。同时，联邦学习对于模型训练的性能与可扩展性要求较高。一方面，需要考虑额外增加的加解密时间开销和密文空间开销以及密文运算和通讯的时间开销。另一方面，随着参与方增加，中心模型的收敛时延、各方之间的通信开销和运算开销将随之增加。因此，如何平衡数据隐私，提高算法收敛速度以及模型训练和各方通信效率等，也是联邦学习乃至机器学习隐私保护技术面临的重要挑战。

4 面向机器学习的隐私保护技术评估指标

为了评估机器学习下隐私保护的性能，相关的评估指标不仅需要涵盖对隐私质量的评估，还要考虑隐私保护技术本身的算法复杂度，对于数据/模型可用性以及可解释性、运算开销等方面的影响。对于原数据共享下的隐私保护技术，本文将适用的隐私性能评估指标分为隐私质量、数据质量与算法复杂度三大类，总结见表 1。对于不分享原数据的隐私保护技术，将从训练损失、模型收敛率、资源消耗、通信开支和时延等 5 个指标维度进行评估。

表 1 原数据共享下的隐私保护性能指标总结

类别	维度	指标	适用技术方法	值域	值域 vs 隐私
隐私质量指标	信息不确定性	匿名集大小	数据限制发布技术	[1，|D|]，|D|为包含某属性的数据总个数	正相关
		隐私熵	通用	[1，|D|]，|D|为包含某属性的数据总个数	正相关
		归一化隐私熵	通用	[0,∞]	正相关
		差分隐私	数据扰动技术、差分隐私等	[0,∞]	负相关
	不可区分度	归一化方差	数据扰动技术、差分隐私等	[0,1]	正相关
		条件隐私熵	通用	[0,∞]	正相关
		置信区间	通用	[0,1]	负相关
		隐藏故障	通用	[0,1]	负相关
数据质量	准确性	区分度量	通用	[0,1]	-
	完整性	损失度量	通用	[0,1]	-
		人工模式	通用	[0,1]	-

	一致性	条件隐私熵*[2]	通用	[0, ∞]	正相关
		K-L 散度*	[0, ∞]	正相关	[0, ∞]
		互信息*	通用	[0, ∞]	负相关
复杂度		算法效率	通用	(0， ∞)	–
		可扩展性	通用	–	–

4.1 原数据隐私保护性能评估指标

（1） 隐私质量

对经过数据限制发布或失真处理后的数据集进行隐私质量评估，根据是否有原数据集进行对比，将其分为信息不确定性与不可区分度两大类指标。

信息不确定性：用来衡量要发布或共享的数据集经过处理后，所增加的信息模糊度。信息不确定性相关的指标往往基于信息论[20]来建立。常用的指标包括匿名集大小[21]，计算为包含同样属性值个体的数目$priv_{ass} \equiv |AS_t|$。隐私熵[22]，用来预测某个变量时的信息不确定性，用变量 X 代表匿名集中某个属性取值的概率分布$p(x)$，隐私熵计算为$priv_{ENT} \equiv \mathrm{H(X)} = -\sum_{x \in X} p(x)\log_2 p(x)$等。由于隐私熵的取值依赖于匿名集中元素的数量，因此隐私熵的绝对值难以公平衡量元素不同的数据集之间的隐私质量之差，因此隐私熵在应用时常被变形为归一化隐私熵[23]。归一化隐私熵可以理解为匿名化数据集所泄露的信息量，计算为$priv_{NE} \equiv \frac{H(X)}{H_0(X)}$，其中，$priv_{MXE} = H_0(X) = \log_2|X|$是 X 可以取得的信息熵的最大值；差分隐私ε[14]，利用ε的值来衡量差分隐私技术实现的隐私等级等。

不可区分度：这类指标在有原数据集或可观测的对比数据集的情况下，来评估能够从不同数据集之间区分出的个体信息或信息差，常用的指标包括：

• 归一化方差[24]等，常用来评估数据失真保护类隐私保护算法中原数据集 X 与扰动后的数据集 Y 之间的分散度，计算为：$priv_{VAR} \equiv \frac{\sigma^2(X-Y)}{\sigma^2(X)}$；

• 条件隐私熵[25]，用来评估在已知扰乱后的数据集中属性 Y 的分布时，对于获取原始数据集中对应的属性 X 分布所需的信息量，$priv_{COE} \equiv \mathrm{H(X|Y)} = 2^{h(X|Y)}$，$\mathrm{h(X|Y)} = -\int_{\Omega_{X,Y}} f_{X,Y}(x,y)\log_2 f_{X|Y=y(x)}dxdy$，其中$f_X(\cdot)$和$f_Y(\cdot)$分别为 X 与 Y 的密度函数；

• 置信区间[26]，用来评估在经过随机化、扰乱类等算法的处理后，原始数据能够从处理后的数据中进行恢复的概率，假设原始数据以$c\%$的概率被估计在$[x_1, x_2]$中，则记为$c\%$置信区间的隐私量为$(x_2 - x_1)$；

• 隐藏故障[3]，定义为在机器学习应用中，进行过隐私保护处理的数据集 Y 中推理出敏感数据量$R_\rho(Y)$与从原始数据集 X 中推理出敏感数据量$R_\rho(X)$相比的比例，计算为：$HF \equiv \frac{R_\rho(Y)}{R_\rho(X)}$。

（2） 数据质量

[2] 加*的指标代表既可以用于评估隐私质量，也可用于评估数据质量。

隐私保护技术对于原数据的处理通常会降低其在机器学习中的可用性。在面向机器学习应用时，变换或处理后的数据可能会降低训练模型的精确度。因此，对于隐私保护技术性能的度量，不仅需要评估数据隐私质量，还需要量化其对数据可用性的影响。在面向机器学习应用时，本文提出 3 个重要的评估维度：准确性、完整性和一致性，相应的指标总结如下。

准确性：用于评估原始数据集与处理后的数据集之间的相似程度，适用于电信网络数据集的指标包括：区分度量（discernibility metric）[27]，用来度量由于泛化导致数据集中有多少记录与给定的记录相同，区分度量越大，信息的损失量越大，准确度越低。

完整性：用来量化原始数据集经过处理后单个数据丢失的程度。在面向机器学习的电信数据集中，除了需要进行脱敏处理的数据，还包含机器学习应用所需的其他数据，可以采用损失度量（misses cost）[28]来评估脱敏后数据集 Y 与原始数据集 X 相比，其中不敏感数据受到脱敏算法影响的数据量，计算为$MC \equiv \frac{R_p(X) - R_p(Y)}{R_p(X)}$，其中$R_p(X)$为原始数据集中不敏感数据元素的数目，$R_p(Y)$为处理后的数据集中不敏感数据元素的数目。人工模式（artifactual patterns）[28] 则用来评价经过处理后的数据集 Y 与原始数据集 X 相比，隐私保护算法带来的无用数据量，计算为$AP \equiv \frac{|Y| - |X \cap Y|}{|Y|}$，其中，$|\cdot|$代表集合的基数。

一致性：用来评估处理前后数据集的正相关性。条件隐私熵，不仅可以用来度量隐私性能，也同时度量了经过处理后的数据集与原数据集相比的信息损失，类似的指标还包括相对信息熵（K-L 散度）[29]，记为ent_{RLE}，用来评估原始的数据集中某个属性分布 p 与处理后的数据集中某个属性分布 q 之间的距离，计算为$ent_{RLE} \equiv D_{KL}(p||q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}$；互信息[30]，用来评估原始数据集 X 经过处理后发布数据集 Y 时的隐私损失量，计算为$priv_{MI} \equiv I(X;Y) = H(X) - H(X \mid Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$，其中，p(x)和 p(y)分别为 X 与 Y 的概率密度分布。当互信息用于评估隐私保护算法的隐私性能时，互信息值越小，隐私级别越高，但同时也说明发布数据集信息量越少。

（3） 复杂度

对于隐私保护类算法，复杂性的度量主要涉及算法的效率和可扩展性[3]。这些指标通常是其他算法所共有的，因此，在本小节只进行简短的介绍。算法效率：为了评估隐私保护算法的效率，一般采用的指标包括对时间例如 CPU 占用时间、计算成本，空间例如算法执行所需的内存量等。而在涉及到数据传输和交换时，带宽的占用率有时也需要考虑在内。可扩展性：用来评估隐私保护算法在不同数据量、不同分布、不同维度等不同特征下数据集的性能保证与适应能力[31]。对于可扩展性的评估，可以通过不断增加数据量或是变换数据分布特征等，衡量算法的效率和性能损失，从而判断隐私保护算法所适用的系统与具体场景。

4.2 不分享原数据的隐私保护性能评估指标

对于不分享原数据的隐私保护技术，所采用的同态加密技术或联邦学习框架，会影响到整个机器学习过程的性能与输出模型的表现，具体包括：

（1）训练损失，用于评估训练后的模型与训练数据的匹配程度，在不分享原数据下的隐私保护技术可能会对机器学习的训练模型的精确度造成一定影响，但训练损失同时也与训练方法的选择有关，因此不分享原数据的隐私保护技术需要结合机器学习模型来进行选择；

（2）模型收敛率，用来评估当多个数据拥有方在本地进行分布式训练时，模型能否收敛至统一的全局模型，以及收敛的效率，它取决于数据拥有方在进行参数更新时所选用的加密技术以及数据拥有方的数据分布是否均衡、梯度更新是否同步等；

（3）资源消耗，数据拥有方为了保护原数据或训练的中间数据，所采用的同态加密或联邦学习框架，会增加对计算和通信资源的消耗，因此需要评估隐私保护下的机器学习所需的内存量、数据传输和交换时带宽的占用率和吞吐量等；

（4）通信开支，基于联邦学习框架的模型训练需要进行本地模型与中心模型之间的参数传输与迭代，这将大大增大训练的时延和带宽的消耗；

（5）时延：包括训练时延、数据传输时延和推理时延，会影响到整个机器学习过程的性能表现与模型的有效度等。

5 机器学习隐私保护发展建议

纵观以上对于机器学习隐私保护技术与评估指标的分析，可以发现现存的隐私保护技术仍较为单一，集中基于差分隐私、同态加密以及联邦学习等框架实现，而在实际部署中数据分布的特点和机器学习任务对于隐私保护算法的设计和应用提出更高和更细致的要求。基于以上分析，针对机器学习下隐私保护技术，本文提出以下发展建议。

（1） 建立完善的机器学习隐私评估体系

本文对机器学习下的隐私保护性能评估指标进行了初步探讨，然而目前还尚未出现完备统一的评估体系。建立完善的评估指标和标准、测评准则与方法等，将为机器学习隐私保护技术的设计提供重要支撑与参考，是保证机器学习隐私必不可少的组成部分。

（2） 设计自适应隐私保护技术框架

开发机器学习模型的隐私保护框架是近年来的一大研究热点。现有的隐私保护技术可扩展性和通用性较差，例如差分隐私的实现需要针对不同的机器学习模型进行设计与调整，隐私保护技术在不同特征的数据集下可实现的保护能力相差较大，因此如何设计一个通用、高效并具备自适应能力的隐私保护框架，构建机器学习隐私保护能力的“底座”，是实现机器学习隐私保护的一大挑战。

（3） 多种技术结合实现机器学习全过程的隐私保护

经过以上分析，可以得出不同的隐私保护技术存在各自的优缺点，联邦学习和同态加密能够保证数据在本地的私有化以及模型数据在参数与计算过程中的机密性和正确性，但会大大增加计算和通信开销；数据失真保护技术的优点在于不会造成过高的计算和通信消耗，可以用于对原数据、模型参数、目标函数等进行扰动以实现差分隐私，一方面这可能会降低模型的可用性与可解释性，另一方面却也可以提高模型的

泛化能力。目前，区块链技术因其具有的去中心化、安全可信和可追溯等特性，被用来为多方参与的模型训练过程提供审计功能[32]。通过探索多种技术结合[33]，来应用于机器学习数据发布、模型训练、数据传输和结果发布等不同阶段，可以弥补单个隐私保护技术的缺陷，实现机器学习全过程的隐私保护。

（4） 在多目标均衡下实现机器学习隐私保护

隐私保护过程与数据的可用性、训练的高效性、模型的收敛度和可解释性以及资源的消耗率等方面存在相互的影响，在机器学习实现与部署过程中，要考虑多个系统指标的平衡，包括数据中包含的个体信息的敏感程度、对模型性能的预期、系统所提供的计算与通信能力等，在多目标均衡以及公平公正的前提下实现机器学习过程的隐私保护。

6 结束语

机器学习技术蓬勃发展的同时，数据治理与隐私保护成为关键影响因素与发展挑战。本文充分调研当前国内外相关法律法规与标准化进展，对机器学习下适用的隐私保护技术进行了归纳分析。为平衡隐私算法对于隐私性、数据质量和算法高效性等方面的相互制约，本文对相应的隐私保护指标进行了总结，探讨了未来机器学习隐私保护的研究方向。总之，机器学习下的隐私保护不断受到更多的关注与重视，在未来将有望实现个性化度量与数据的按需保护，将实现多种隐私保护机制的联合使用与自适应选择，并将针对机器学习中的隐私泄露攻击出现越来越可靠的防御手段。

参考文献：

[1] LANGHEINRICH M. Privacy in ubiquitous computing [Z]. Ubiquitous Computing Fundamentals, 2009(3): 95–159.

[2] United Nation General Assembly.Universal Declaration of Human Rights. [EB]. 2020.

[3] Bertino E, Lin D, Jiang W. A survey of quantification of privacy preserving data mining algorithms[M] Privacy-preserving data mining, Springer, 2008, pp.183-205.

[4] 宋蕾, 马春光, 段广晗.机器学习安全及隐私保护研究进展[J]. 网络与信息安全学报, 2018, 4(8):5:1-11.

SONG L, MA C, DUAN G. Machine learning security and privacy: a survey [J]. Chinese Journal of Network and Information Security, 2018, 4(8): 1-11.

[5] 赵镇东,常晓林,王逸翔.机器学习中的隐私保护综述[J].信息安全学报，2019, 4(5): 1-13.

ZHAO Z, CHANG X, WANG Y. A Survey of Privacy Preserving in Machine Learning[J]. Journal of Cyber Security, 2019, 4(5): 1-13.

[6] ROUANI B D, SAMRAGH M, JACIDI T, et al. Safe machine learning and defeating adversarial attacks[J]. IEEE Security & Privacy, 2019, 17(2): 31-38.

[7] AL-RUBAIE M, CHANG J M. Privacy-preserving machine learning: Threats and solutions[J]. IEEE Security & Privacy, 2019, 17(2): 49-58.

[8] 刘俊旭, 孟小峰. 机器学习的隐私保护研究综述[J]. 计算机研究与发展, 2020, 57(2): 346-362.

LIU J, MENG X. Survey on privacy-preserving machine learning[J]. Journal of Computer Research and Development, 2020, 57(2): 346-362.

[9] SWEENEY L. K-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and

Knowledge-Based Systems, 2002, 10(5): 557-570.
[10] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. L-diversity: privacy beyond k-anonymity[J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007, 1(1): 3.
[11] LI N, LI T, VENKATASUBRAMANIAN S. T-closeness: privacy beyond k-anonymity and l-diversity[C]//Proceedings of 2007 IEEE 23rd International Conference on Data Engineering.Piscataway: IEEE Press,2007: 106-115.
[12] DWORK C, MCSHERRY F, NIAIIM K, et al. Calibrating noise to sensitivity in private data analysis[C]// Theory of Cryptography Conference. Berlin: Springer, 2006.
[13] DWORK C, SMITH A, STEINKE T, et al. Exposed! a survey of attacks on private data[J]. Annual Review of Statistics and Its Application, 2017: 61-84. .
[14] XU C, REN J, ZHANG D, et al. GANobfuscator: mitigating information leakage under GAN via differential privacy[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(9): 2358-2371.
[15] PHAN N H, WANG Y, WU X, et al. Differential privacy preservation for deep auto-encoders: an application of human behavior prediction[C]//Proceedings of Thirtieth AAAI Conference on Artificial Intelligence, California: AAAI Press, 2016.
[16] JAYARAMAN B, EVANS D. When relaxations go bad:" differentially-private" machine learning[J]. arXiv:1902.08874, 2019.
[17] RIVEST R L, ADLEMAN L, DERTOUZOS M L. On data banks and privacy homomorphisms[J]. Foundations of secure computation, 1978, 4(11): 169-180.
[18] LI T, SAHU A K, TALWALKAR A, et al. Federated learning: challenges, methods, and future directions[J]. arXiv:1908.07873, 2019.
[19] NASR M, SHOKRI R, HOUMANSADDR A. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks[J]. arXiv :1812.00910, 2018.
[20] YAO A C. Protocols for secure computations[C]//Proceedings of 23rd Annual Symposium on Foundations of Computer Science . Piscataway: IEEE Press, 1982: 160-164.
[21] KESDOGAN D, EGNER J, BUSCHKES R. Stop-and-go-MIXes providing probabilistic anonymity in an open system[C]//Prceedings of International Workshop on Information Hiding, Berlin: Springer, ,1998: 83-98.
[22] SERIGANTOV A, DANEZIS G. Towards an information theoretic metric for anonymity[C]//Proceedings of International Workshop on Privacy Enhancing Technologies.Berlin: Springer, , 2002: 41-53.
[23]DIAZ C,. Towards measuring anonymity[C]//Proceedings of International Workshop on Privacy Enhancing Technologies, Berlin, :Springer, , 2002.
[24]OLIVEIRA S R M, ZAIANE O R. Privacy preserving clustering by data transformation[J]. Journal of Information and Data Management, 2010, 1(1): 37-37.
[25] DIAZ C, TRONCOSO C, DANEZIS G. Does additional information always reduce anonymity?[C]//Proceedings of the 2007 ACM workshop on Privacy in electronic society,New York:ACM Press,2007: 72-75.
[26] AGRAWAL R, SRIKANT R. Privacy-preserving data mining[C]//Proceedings of the 2000 ACM SIGMOD international conference on Management of data, New York:ACM Press,2000: 439-450.
[27] BAYARDO R J, AGRAWAL R. Data privacy through optimal k-anonymization[C]//21st International conference on data engineering (ICDE'05).Piscataway: IEEE Press, 2005: 217-228.
[28] OLIVEIRA S R M, ZAIANE O R. Privacy preserving frequent itemset mining[C]//Proceedings of the IEEE International Conference on Privacy, Security and Data Mining-Volume 14,[S.l:s.n.], 2002: 43-54.
[29] DENG Y, PANG J, WU P. Measuring anonymity with relative entropy[C]//International Workshop on Formal Aspects in

Security and Trust. Berlin:Springer,2006: 65-79.

[30] LIN Z, HEWETT M, ALTMAN R B. Using binning to maintain confidentiality of medical data[C]//Proceedings of the AMIA Symposium,[S.l.:s.n.], 2002: 454.

[31] BONDI A B. Characteristics of scalability and their impact on performance[C]//Proceedings of the 2nd International Workshop on Software and Performance,[S.l:s.n.],2000: 195-203.

[32] LU Y, HUANG X, DAI Y, et al. Blockchain and federated learning for privacy-preserved data sharing in industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2019, 16(6): 4177-4186.

[33] KANG J, XIONG Z, NIYATO D, et al. Reliable federated learning for mobile networks[J]. IEEE Wireless Communications, 2020, 27(2): 72-80.

[作者简介]

刘姿杉（1992－ ），女，博士，中国信息通信研究院工程师，主要研究方向为人工智能、宽带接入、数据安全隐私等。

程强（1977－ ），男，中国信息通信研究院高级工程师，主要研究方向为人工智能、宽带接入、家庭网络等。

吕博（1981－ ），男，博士，中国信息通信研究院高级工程师，主要研究方向为人工智能、量子计算、高精度时间同步等。