机器学习安全及隐私保护研究进展

宋蕾，马春光，段广晗
（哈尔滨工程大学计算机科学与技术学院，黑龙江 哈尔滨 150001）

摘　要：机器学习作为实现人工智能的一种重要方法，在数据挖掘、计算机视觉、自然语言处理等领域得到广泛应用。随着机器学习应用的普及发展，其安全与隐私问题受到越来越多的关注。首先结合机器学习的一般过程，对敌手模型进行了描述。然后总结了机器学习常见的安全威胁，如投毒攻击、对抗攻击、询问攻击等，以及应对的防御方法，如正则化、对抗训练、防御精馏等。接着对机器学习常见的隐私威胁，如训练数据窃取、逆向攻击、成员推理攻击等进行了总结，并给出了相应的隐私保护技术，如同态加密、差分隐私。最后给出了亟待解决的问题和发展方向。

关键词：机器学习；安全威胁；防御技术；隐私保护
中图分类号：TP309.2
文献标识码：A
doi: 10.11959/j.issn.2096-109x.2018067

Machine learning security and privacy: a survey

SONG Lei, MA Chunguang, DUAN Guanghan
School of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China

Abstract: As an important method to implement artificial intelligence, machine learning technology is widely used in data mining, computer vision, natural language processing and other fields. With the development of machine learning, it brings amount of security and privacy issues which are getting more and more attention. Firstly, the adversary model was described according to machine learning. Secondly, the common security threats in machine learning was summarized, such as poisoning attacks, adversarial attacks, oracle attacks, and major defense methods such as regularization, adversarial training, and defense distillation. Then, privacy issues such were summarized as stealing training data, reverse attacks, and membership tests, as well as privacy protection technologies such as differential privacy and homomorphic encryption. Finally, the urgent problems and development direction were given in this field.

Key words: machine learning, security threats, defense technology, privacy

1　引言

机器学习是人工智能技术之一，近些年来，随着其不断成熟而飞速发展，大量企业在机器学习领域取得了突破性进展，如在医疗、图像处理、网络空间安全等领域中都得到了广泛应用。随着深度学习的兴起，机器学习又迎来新的一波发展热潮，推进人工智能向前迈进一大步。在机器学习火速发展的同时，其安全与隐私问题也引起了人们的关注。机器学习的安全和隐私的威胁已经

收稿日期：2018-05-07；**修回日期：**2018-07-02
通信作者：马春光，machunguang@hrbeu.edu.cn
基金项目：国家自然科学基金资助项目（No.61472097）
Foundation Item: The National Natural Science Foundation of China (No.61472097)

阻碍机器学习及人工智能的发展。在自动驾驶、财政系统、健康等系统中，机器学习的安全性问题威胁人们的切身利益甚至健康生活。在云计算服务中的机器学习即服务（MLaaS），人们可能考虑数据泄露的问题而放弃便捷云服务。因此，如何保障机器学习的安全及如何保护用户隐私成为机器学习发展的基础，人们就此问题展开研究。虽然此项研究刚刚起步，但已经取得了一定进展。

在机器学习中，安全是保障用户数据被正确地使用，保障机器学习的可用性和完整性。安全和隐私是 2 个不同但紧密相关的概念，安全是保护用户隐私的基础。隐私权是自 1948 年以来联合国《世界人权宣言》中所包含的一项基本人权，隐私的法律意义为不愿告人或不便告人的事情，和别人无关，关于自己利益的事。隐私是一个没有公认标准定义的复杂概念[1]。在机器学习中隐私被定义为人们有权利决定自己的私有数据不被公开。

2　机器学习及其敌手模型

2.1　机器学习

机器学习通过计算手段利用经验改善系统的自身性能。经验即数据，计算机系统利用现有数据进行学习，产生模型进而对未来的行为做出决策判断。机器学习解决问题的过程如图 1 所示。

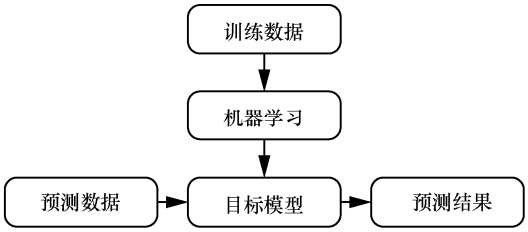


图 1　机器学习解决问题的过程

机器学习解决问题过程分为训练和预测 2 个阶段。机器学习的学习方式主要分为：有监督学习、无监督学习和强化学习。有监督学习的训练数据集有标签，输入数据可以形式化表示为 $(x,\bar{y})$。有监督学习多用于分类、回归问题，如图像识别、垃圾邮件分类等。无监督学习训练数据集不带标签，输入数据为(x)。无监督学习多用于聚类、数据降维等，如预训练、入侵检测等。强化学习通过试探与环境交互以获得策略的改进。强化学习是机器学习的一个重要分支，第一个击败人类职业围棋选手的人工智能程序 AlphaGo[2] 就是使用有监督和有无监督相结合的强化学习技术。在训练阶段后获得目标模型 $h_\theta(x)$，人们可以利用目标模型进行预测做出决策。

为了便于描述，本文使用机器学习进行的分类任务为例，描述在训练与预测阶段中可能遇到的安全与隐私问题。机器学习其他应用面临的安全与隐私问题与其类似。

2.2　敌手模型描述

在描述机器学习威胁时需要考虑攻击者的能力、目标等，这里的攻击者称为敌手。敌手模型可以从敌手目标、敌手知识、敌手能力、敌手策略 4 个维度刻画[3]。根据敌手模型的不同产生的威胁也不同。

敌手目标：敌手目标分为破坏机器学习的机密性、完整性、可用性。机密性是指包含用户隐私的敏感信息不被泄露。敌手不仅可以窃取含有敏感信息的训练数据，还可以通过暴露目标模型信息及其预测结果达到其目的，严重威胁用户隐私。完整性和可用性是对于机器学习模型输出而言，均影响用户正常使用模型进行预测，但侧重点不同。完整性威胁指敌手诱导模型行为或者使模型在预测中输出指定分类标签。可用性威胁指阻止用户获得模型正确的输出或者阻止获取模型本身的一些特性，使其在目标环境下不可信赖。完整性威胁和可用性威胁目标相似，敌手采取的手段也相近。敌手在攻击过程中的目标往往不是单一的，如在威胁机密性下获取目标模型，通过分析目标模型得出模型的脆弱性对其发起攻击，对模型可用性和完整性产生威胁。

敌手知识：敌手知识包括模型的训练数据及特征、模型结构及参数、决策函数、访问目标模型得到反馈信息等。根据敌手已知信息的多少，敌手知识分为有限知识和完全知识。在预测阶段中，根据敌手知识可将攻击分为白盒攻击和黑盒攻击。白盒攻击指敌手得到目标模型，了解模型的内部结构及其参数。黑盒攻击指敌手只能访问目标模型，根据输入数据得到模型输出相应的预测结果。在实际中，白盒攻击不易实现，黑盒攻击更为常见。

敌手能力：敌手能力指敌手在具有一定知识

背景下，对模型或者训练数据、测试数据的控制能力。根据控制能力的不同，可以把敌手分为强敌手和弱敌手。在训练阶段中，敌手访问训练数据、注入恶意数据、直接修改数据，这些敌手能力是逐渐增强的。在预测阶段，白盒攻击中的敌手能力比黑盒攻击强。如果敌手可以改变目标模型的学习策略，则防御者很难在预测阶段抵抗这种逻辑上的改变所带来的攻击。

敌手策略：敌手策略指敌手为达到攻击目标，根据自身的知识和能力，采取的具体攻击方式，如修改数据集标签信息、注入恶意数据、逆向攻击提取敏感数据等。

3　机器学习安全威胁及防御技术

3.1　机器学习常见的安全威胁

在人们的日常生活中，机器学习越来越普及。机器学习的安全威胁严重影响人们的正常生活甚至生命健康。例如，在汽车自动驾驶过程中，敌手可以通过伪造交通标志影响汽车的正常驾驶，易导致发生车祸；在垃圾邮件检测中，敌手可以通过引入积极的词汇逃避垃圾邮件的分类。表 1 给出了常见的安全威胁，这些安全威胁分别针对机器学习过程中的训练阶段和预测阶段，如图 2 所示。

3.1.1　训练阶段的安全威胁

在训练阶段，机器学习使用训练数据集训练目标模型，通过学习数据的内在特征以得到决策假设函数，在预测阶段则应用目标模型进行预测。目标模型预测的有效性依赖于训练数据集和预测数据集属于同一分布。例如，预测数据与训练数据分布不同，则在实际预测阶段会出现偏差。因此，很多敌手会通过修改训练数据的分布对目标模型发起攻击。

投毒攻击（poisoning attack）[3]是在训练阶段中常见的攻击手段，敌手对训练数据进行修改、删除或注入精心制作的恶意数据，改变训练数据原有的分布，使学习算法在逻辑上发生改变进而威胁目标模型。一旦模型训练成功后，在预测阶段这种威胁是很难防御的。文献[4]表明，当模型预测误差小于 ε 时，其最大容忍修改训练数据集的概率是 b，b 应满足 $b \leqslant \frac{\varepsilon}{\varepsilon+1}$。

敌手修改训练数据时，可以直接修改数据标签，将训练数据 $(x,\bar{y})$ 修改成 $(x,\bar{y}')$，使模型在学习过程中学习错误的数据特征。文献[5]中敌手随机修改 40%的训练数据标签，使模型在二分类任务中无法进行正常分类。

除了直接修改数据标签外，敌手可以通过注入精心制造的恶意数据实现投毒攻击，如文献[6-7]。敌手通过注入精心制作的恶意样本改变训练样本的分布，训练后模型的决策边界发生变化，使预测阶段模型的精度降低，甚至敌手可以诱导模型输出指定的错误分类标签。投毒攻击的攻击过程如图 3 所示。

图 3(a)中实线表示对二分类样本进行训练后的分类器，虚线表示在训练阶段中加入恶意样本后，被破坏的分类器。图 3(b)表示在预测阶段，因投毒攻击对分类器产生了影响，导致部分测试样本被错误分类。

表 1　　**机器学习中常见的安全威胁**

阶段	敌手策略	敌手目标	敌手能力	敌手知识
训练阶段	投毒攻击	完整性/可用性	修改训练数据	有限知识
预测阶段	对抗攻击	完整性/可用性	制作对抗样本	黑盒/白盒
	询问攻击	完整性/可用性	访问目标模型	黑盒

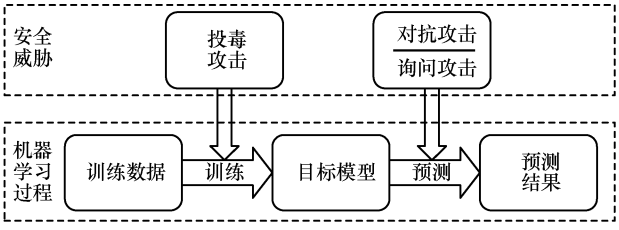
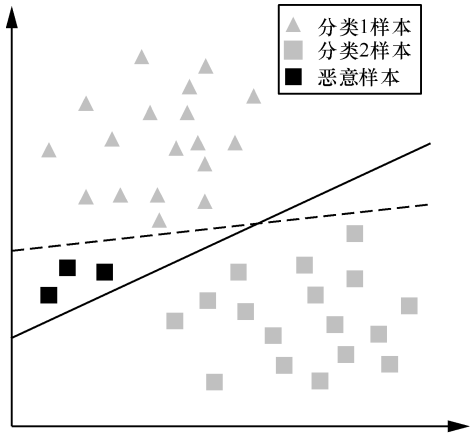
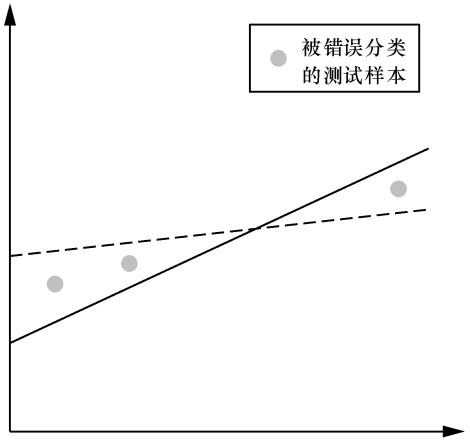


图 2　机器学习常见的安全性威胁



(a) 训练阶段注入恶意样本



(b) 预测阶段分类器做出错误判断

图3　投毒攻击的攻击过程

训练阶段往往都是集中、离线和封闭的。敌手不易获取训练数据，不能直接在原始训练数据中投毒。而在一些在线学习或者需要定期更新模型的训练中，需要不断加入新的训练数据，这就给敌手可乘之机，在新收集的数据中投毒，如文献[8-9]均采取此类做法。对防御者而言，当新收集的数据集分布发生明显变化时，需提高警惕。

3.1.2　预测阶段的安全威胁

经过训练后，得到机器学习目标模型，部署好后进行推理和预测。在预测阶段中，敌手无法修改训练数据集，但可以访问目标模型获得有效的信息对模型发起攻击，使模型在预测阶段发生错误。2014年，Szegedy等[10]首先发现对图片添加轻微扰动可以欺骗神经网络，此过程称对抗攻击（adversarial attack），敌手精心制造使模型错分类的样本称为对抗样本（adversarial example）。根据敌手知识，将此阶段的攻击分为白盒攻击和黑盒攻击。

白盒攻击：敌手已知目标模型的结构和参数，虽然在实际中不易实现，一旦敌手获取目标模型则对机器学习造成极大的威胁。敌手可以通过分析目标模型结构来构造对抗样本进行对抗攻击。Szegedy等[10]将对抗样本作为输入，在预测阶段使目标模型分类错误。他们将搜索对抗样本形式化表示为以下优化问题。

$$\arg\min_r h(x+r)=l \text{ s.t } x+r\in D \tag{1}$$

其中，x为输入，h为目标模型的分类函数，r为扰动。对抗样本$x^*=x+r$，$x+r$仍然在原输入分布D内，但是模型预测时将x^*错分到l标签下，这种将对抗样本错分到指定分类类别为针对目标攻击（targeted attack）。与之相反，分类器以较高的置信度输出错误的分类为非针对目标攻击（non-targeted attack）[11]。为了求解方程(1)，Szegedy使用L-BFGS优化算法得到最小扰动r，使扰动后的图像可以欺骗深度学习模型，使之分类错误。

Goodfellow等[12]使用FGSM（fast gradient sign method）解决上述优化问题，利用线性假设（现代DNN模型采用ReLU作为激活函数而不再sigmod，鼓励线性计算同时也易受对抗攻击）可以快速搜索对抗样本。

$$x^*=x+\in \text{sign}\left(\nabla_x J_h(\theta,x,y)\right) \tag{2}$$

其中，J为目标模型的代价函数，sign(·)为方程的符号，$\in$为限制扰动范围的标量。尽管是基于线性假设的近似求解，也使目标模型在MNIST数据集上达到89.4%错分类。同时，文献[12]首次揭示对抗样本的分布特征，即对抗样本大多存在模型的决策边界附近。2017年，Kurakin等[13]在FGSM基础上提出BIM（Basic Iterative Method）快速生成对抗样本方法，在每次迭代中使用FGSM生成的对抗样本。此后文献[14-15]均在此基础上进行研究，取得不错效果。

Moosavi-Dezfooli等[16]提出Deepfool方法，通过迭代计算的方法生成最小规范对抗扰动，直到找到距离正常样本x最接近的决策边界，跨越边界找到扰动最小的对抗样本。作者证明Deepfool生产的对抗样本在相似的欺骗率下扰动比FGSM小。

Papernot等[17]提出JSMA（jacobian-based saliency map），该方法限制扰动L_0范数而不是L_2

或者 L_∞ 范数，意味着只需要修改图片的几个像素而不是扰乱整张图片就可以欺骗分类器。在极端情况下，Su 等[18]只修改图像中的一个像素值就可以实现对抗攻击。

黑盒攻击：敌手得不到目标模型的内部结构和参数，但可以通过 API 进行访问目标模型，尤其在云环境下的机器学习即服务更容易访问 API 接口进行黑盒攻击。询问攻击（Oracle attack）通过观察特定的输入对应的输出信息，建立与目标模型相似的模型进行攻击，Oracle 攻击的有效性与询问目标模型的输入及查询次数密切相关。Lowd 等[19]评估了使模型错分类需要询问次数的成本。

Szegedy 等首先观察到对抗样本转移性，即被一个模型错分类的对抗样本对其他模型也适用。Moosavi-Dezfooli 等[20]表明不同模型中存在通用扰动（universal perturbation），使用通用扰动产生对抗样本可以在 ImageNet 训练的目标模型上传递。Papernot 等[21]利用对抗样本跨模型传递性（cross-model transferability）实现黑盒攻击，该攻击利用敌手产生的合成输入训练一个替代模型（substitute mode），利用替代模型制作对抗样本，这些对抗本可以被原本的目标模型错误分类。作者表明，利用 MetaMind 在线训练 DNN 模型错分类可以达到 84.24%，之后作者使用亚马逊云服务实现逻辑回归时错分类可以达到 96%[22]。

表 2 描述机器学习预测阶段常见的安全威胁，并对敌手知识和攻击目标进行对比分析，可知黑盒攻击敌手得不到目标模型的结构和参数信息，更难从中提取模型的决策边界信息，相较于白盒，攻击能力弱一些，不易实现针对目标攻击，但在现实中更具有普遍性。

表 2　机器学习预测阶段安全威胁对比分析

攻击类型	访问模型输出	获取模型内部结构	针对目标攻击	非针对目标攻击
白盒攻击	√	√	L-BFGS[10]、FGSM[12]、JSMA[17]	BIM[13]、Deepfool[16]
黑盒攻击	√			文献[20-22]

3.2　机器学习安全性防御技术

机器学习的安全性问题威胁系统的可用性和完整性，敌手多依赖于测试数据与训练数据分布不同发起攻击。提高目标模型的顽健性可以对预测阶段中出现的未知样本有良好的适应性，在出现恶意样本时模型也可以正常进行预测。

正则化（regularization）。通过为代价函数添加正则项（也叫惩罚项）提高目标模型的泛化能力，在预测中遇见未知数据集具有良好的适应性抵抗攻击。Barreno 等[23]提出使用正则化方法保障机器学习安全的建议。Biggio 等[5]利用正则化方法限制训练 SVM 模型时敌手修改训练标签的脆弱性。Gu 等[23]使用更平滑惩罚项训练深度收缩网络（deep contractive network）抵御攻击。文献[24-26]使用正则化方法提高算法的顽健性，在抵抗攻击时取得良好的效果。

对抗训练（adversarial training）。在训练数据集中引入对抗样本，通过合法化的对抗样本对目标模型的训练提供模型的顽健性。合法化的对抗样本模仿在预测阶段可能的数据分布，Szegedy 等[10]首先通过注入对抗样本，修正其标签使模型面对敌手时更具顽健性。Goodfellow 等[27]利用对抗训练将在 MNIST 数据集上的错误识别率从 89.4%降低到 17.9%。Huang 等[28]通过惩罚错分类的对抗样本增加模型的顽健性。Tramèr 等[29]提出联合对抗训练（ensemble adversarial training）增加对抗样本多样性，但在对抗训练中引入所有未知攻击的对抗样本是不现实的，对抗训练的非适应性导致对抗训练的局限性。

防御精馏（defensive distillation）。2016 年，Papernot 等[30]在 distillation[31]技术的基础上提出防御精馏技术抵抗攻击。原 distillation 技术旨在将大规模模型压缩为小规模并保留原有的准确性，而 defensive distillation 不改变模型规模的大小，产生输出表面更平滑的、对扰动不敏感的模型提高模型的顽健性，作者表示，使用 defensive distillation 方法可以将对抗样本攻击的成功率降低 90%。然而，在黑盒攻击中并不能确保 defensive distillation 的有效性，因此，2017 年，Papernot 等[32]提出可扩展的防御精馏技术。

防御者在防御时与敌手做博弈游戏，防御者制定防御策略，敌手做出最佳响应[33]。防御者代价函数为

$$E_{x,y} \sim D[l_h(\Delta(x), y)] \tag{3}$$

敌手代价函数可为

$$E_{x,y} \sim D[l_h^a(\Delta(x), y) + c(x, \Delta(x))] \tag{4}$$

其中，$\Delta(\cdot)$ 代表敌手修改样本的函数，$c(x, \Delta(x))$ 为修改样本的代价。防御者与敌手之间的博弈过程可表示为敌手最小化制造样本的成本，防御者最小化存在对抗样本时的代价函数，如式(5)所示[34]。防御者与敌手不断交互博弈的过程增加目标模型的顽健性[35]。

$$\min \sum_{i=1}^{n} E_{x,y} \sim D[l_h(\Delta(x), y)] \tag{5}$$

$$\Delta \in \arg\min_{\Delta} \sum_{i=1}^{n} E_{x,y} \sim D[l_h^a(\Delta(x), y) + c(x, \Delta(x))]$$

除了提高模型的顽健性外，直接拒绝恶意样本也是提高模型安全性的重要手段。可以利用数据清洗直接将投毒数据去除来防御投毒攻击。同样，也可以直接丢弃被检测判定为对抗样本的数据来抵御对抗攻击。文献[36-37]先检测预测样本是否为合法样本，若为合法样本则进行预测，若为对抗样本则直接丢弃，在抵御对抗攻击取得一定效果。

4　机器学习隐私威胁及隐私保护技术

4.1　机器学习常见的隐私威胁

机器学习模型会无意识记录一些训练数据，而一些训练数据涉及人们的隐私，如习惯、爱好、地理位置等。在机器学习的健康系统中，隐私威胁不仅泄露病人隐私，敌手可能发起攻击修改病人的用药剂量导致病人生命危险。表 3 给出了常见的隐私威胁，这些隐私威胁分别针对机器学习过程中的训练阶段和预测阶段，如图 4 所示。

表 3　　机器学习中常见的隐私威胁

阶段	敌手策略	敌手目标	敌手能力	敌手知识
训练阶段	窃取训练数据	机密性	获得训练数据	有限知识
预测阶段	逆向攻击	机密性	提取模型/训练数据信息	黑盒/白盒
	成员推理攻击	机密性	访问目标模型	黑盒

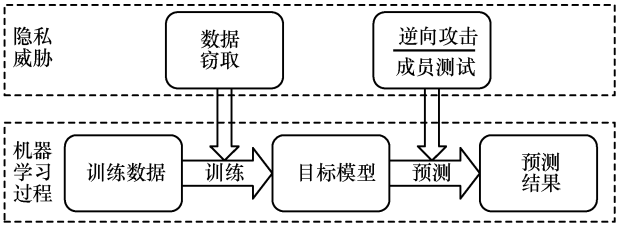


图 4　机器学习中常见的隐私威胁

4.1.1　训练阶段的隐私威胁

机器学习训练方式分为集中式和联合分布式，如图 5 所示。

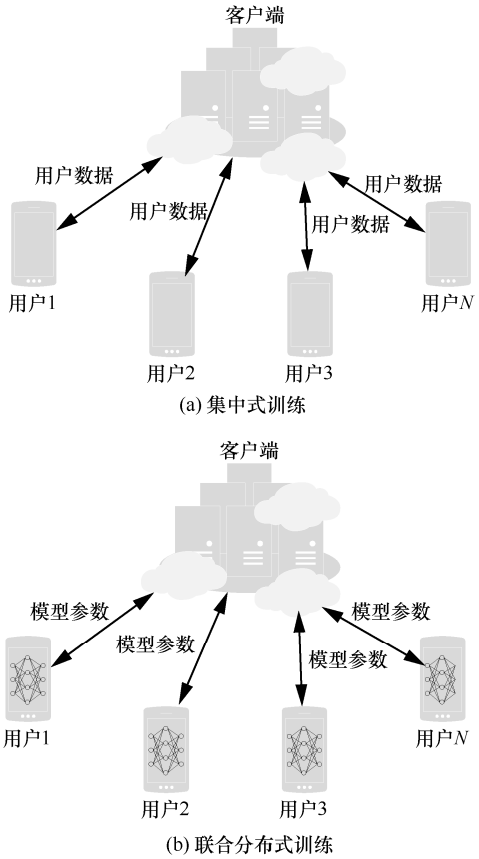


图 5　机器学习训练方式

大型公司多用集中训练方式，因为拥有足够多的用户便于收集大量的训练数据。目前，针对公司收集用户数据保护用户隐私业界没有一个统一的衡量标准。在收集用户数据过程中，会暴露一些用户的隐私，Google 和 Apple 公司采用差分隐私的方式保护用户数据，在用户数据中，加入噪声，使单一的数据没有现实意义，而统计信息具有应用价值。为了扩大训练数据集得到精确的目标模型，一些数据提供方需要进行协作“分享”数据，共同训练目标模型。“分享”不是直接对其他参与方公开数据，各个参与方独自在各自数据集上训练自己的模型，与其他参与方共享训练结果，从而间接分享各自的训练数据。联合训练模型易受不诚实的参与方攻击，恶意窃取其他参与者数据。Hitaj 等[38]利用生成对抗网络（GAN, generative adversarial networks），对联合分布式训练深度学习模型发起攻击，任意参加训练的用户

都可能成为敌手，生成与其他参与者训练数据无限逼近的假样本来窃取他人隐私。

4.1.2　预测阶段的隐私威胁

提取训练数据信息：在预测阶段中通过逆向攻击可以提取训练数据（部分或全部）或训练数据的统计特征。2014 年，Fredrikson 等[39]在药物剂量预测任务中，结合病人的人口属性统计信息可以恢复患者基因组信息。2015 年，Fredrikson 等[40]通过观察目标模型预测结果，反过来重建模型训练时使用的人脸图像数据。Ateniese 等[41]分析模型确定其训练数据是否符合某种统计特征。2017 年，Shokri 等[42]提出成员推理攻击（membership inference attack），给定一条记录可以判定是否在训练数据集中。

提取目标模型信息：Tramer 等[43]表示可以利用观察输入–输出对提取出模型信息。对于一个 N 维的线性模型，理论上通过 N+1 次查询就能窃取到这个模型。可以简单地归结为已知 $(x;h_\theta(x))$ 求解 θ。同时可以利用询问得到的输入–输出对训练辅助模型来模拟目标模型，通过分析辅助模型寻找目标模型的脆弱性对其发起攻击。

表 4 描述机器学习中常见隐私威胁对比分析，可见对模型的恶意访问和暴露模型参数会泄露用户隐私，合理限制目标模型的访问可以有效防止隐私泄露。

4.2　机器学习隐私保护技术

用于训练机器学习的数据包含大量用户隐私，如照片、地理位置、医疗信息等。目前，常用于机器学习的隐私保护技术有同态加密技术和差分隐私技术。

4.2.1　基于同态加密的机器学习隐私保护技术

同态加密[44]（homomorphic encryption）允许用户直接在密文上做运算，得到的结果解密后与在明文下运算结果一致，是最直接有效保护用户隐私的一项技术。在云存储中，云端的目标模型可以直接对密文进行预测。但同态加密运算只支持加法和乘法运算的多项式运算，而机器学习过程中包含幂运算（如 sigmod 激活函数），研究人员就此问题进行研究。

使用同态加密技术对机器学习进行隐私保护，可以在预测阶段中对加密数据直接进行预测，预测结果也是密文，将结果返回给用户自行解密来保护用户数据隐私。2016 年，Dowlin 等[45]提出 CryptoNets 神经网络模型可直接作用于密文做预测。作者假设在云端已经有应用明文训练好的神经网络模型，通过变换使目标模型用于密文预测，利用低阶多项式 $\mathrm{sqr}(z):=z^2$ 代替激活函数。虽然是近似模拟神经网络模型，但用卷积神经网络（CNN）在 MNIST 数据集上识别精度达到 99%。2017 年，Hesamifard 等[46]利用 Chebyshef 多项式近似模拟激活函数，与 CryptoNets[45]不同的是，作者在训练阶段将激活函数替换成低阶多项式，相较 CryptoNets 在 MNIST 数据集上精度提高 0.52%。

为了解决基于全同态加密技术的机器学习计算开销大问题，Baryalai 等[47]利用 Paillier 加法同态技术减少运算开销，提高运算速度。作者提出非共谋双云模型（Cloud A，Cloud B），神经网络模型部署在 CloudA 上，当用户将加密后的数据请求 Cloud A 做出预测时，Cloud A 在密文上求出加权和 $sum=\omega x+b(\omega:weight,b:basis)$，将 sum 密文发送给 Cloud B，用户和 Cloud B 互通密钥进行解密，然后计算激活函数等非多项式运算，将结果再次加密后发送给 Cloud A，依次迭代。作者虽然做了安全性分析，但在实际操作中并未考虑 Cloud A 与 Cloud B 之间的同步问题。

以上方法都是在预测阶段保护用户数据，理论上在训练阶段也可用加密技术保护敏感数据集。Xie 等[48]利用 Stone-Weierstrass 理论[49]。

$$\sup_{x\in X}\|N(x)-P(x)\|<\in \tag{6}$$

其中，N 为连续函数，P 为多项式，$\in$>0，得出

表 4　　**机器学习中常见的隐私威胁对比分析**

隐私威胁	训练阶段	预测阶段	访问模型输出	访问模型内部结构	提取模型结构信息	提取训练数据信息
文献[38]	√		√	√		√
文献[41]		√	√	√		√
文献[43]		√	√		√	
文献[39,40,42]		√	√			√

$$\sup_{x\in X} \| N(x) - D(N'(E(x))) \| < \in \quad (7)$$

其中，N 为神经网络，E 为加密函数，D 为解密函数，$\in>0$，神经网络模型均能用多项式近似模拟，提出 crypto-nets 在密文上做预测。同时，作者就直接使用密文训练神经网络，在几种场景下的适用性做出讨论，得出：1)只有在样本数量很小，或者浅层网络中才能适用；2)不同数据集所有者共同训练模型时，数据提供者利用各自密钥加密数据并不可行，利用相同密钥又不能保护用户敏感数据，这种情况下适用安全多方计算（MPC）；3)已存在用明文训练好的目标模型，让此目标模型适用于自身敏感数据时，可用自身敏感数据对 crypto-nets 网络进行微调。Zhang 等[50]给出直接在密文上训练神经网络的解决方案，利用 Taylor 公式实现对激活函数等非多项式函数的模拟，用户将全同态加密复杂运算交给云端处理，每次反向传播过程后，用户从云服务器下载网络参数（*weight*, *basis*）进行解密，然后重新加密后上传到云端进行迭代运算。这样可以避免电路深度加深，无法正确解密问题。

目前，利用加密技术保护机器学习用户敏感数据多用于预测阶段，虽然在训练阶段中理论上可行，但是深度学习本来就大量消耗计算资源，加密技术计算开销大，网络训练过程慢。如将复杂计算交给云端处理，需注意神经网络模型训练过程是一个迭代的过程，随着运算次数的增多，同态计算电路深度加深，一旦超过阈值，将无法解密得到正确的计算结果。同时也要考虑客户端和云端通信等问题。

4.2.2 基于差分隐私的机器学习隐私保护技术

差分隐私（differential privacy）是一种被广泛认可的严格的隐私保护技术。2006 年，微软的 Dwork[51]提出差分隐私概念，通过引入噪声使至多相差 1 个数据的 2 个数据集查询结果概率不可分。具体定义如下。

$$\Pr[A(T) \in S] \leqslant e^{\varepsilon} \Pr[A(T') \in S] + \delta \quad (8)$$

其中，T、T' 为相差至多一条数据的相邻数据集合。ε 为隐私预算，ε 越小，提供隐私保护能力越强。δ 代表可容忍的隐私预算不成立的概率。防御者可以在训练阶段和预测阶段分别在训练数据集和预测结果中引入噪声来保护隐私，但噪声的引入会降低模型的精度。随着访问次数的增多，噪声随之增加使模型的可用度降低。如何平衡保护隐私和目标模型的精度是研究人员的重点关注方向。下面分别介绍在集中式学习和联合分布式学习中，研究人员如何使用差分隐私技术保护用户隐私。

在集中式学习中，为保护敏感训练数据，2016 年，Abadi 等[52]提出基于差分隐私的深度学习算法，利用随机梯度下降过程中对梯度增加扰动来保护训练敏感数据，并在差分隐私框架下对隐私成本进行了精细的分析，经实验证明，可以在隐私成本可控的情况下完成深层网络训练。2017 年，Papernot 等[53]提出用半监督知识迁移解决深度学习中训练数据隐私泄露问题，通过教师模型全体的隐私聚合（PATE, private aggregation of teacher ensembles）为训练数据提供通用的强健隐私保护。作者将敏感数据分割成 N 个不相交的数据集，在每个数据集上独立训练得到 N 个教师模型，通过对投票引入噪声，算出得票最高的预测结果。然后用公开不含标签的数据集（利用教师模型标注）训练学生模型进行知识迁移传递，最终部署学生模型进行预测服务。不使用敏感数据直接训练公开模型，防止逆向攻击提取原始敏感数据。Beaulieujones 等[54]利用文献[52]中方法，提出差分隐私辅助分类生成对抗网络生成医疗临床数据。郭鹏等[55]对文献[54]提出改进，实现对生成对抗网络进行差分隐私保护。

在联合分布式学习中，数据所有者共同学习具有同一目标的网络模型，在各自数据集上独立训练却共享学习结果。2015 年，Shokri 和 Shmatikov[56]首次将隐私保护概念引入深度学习中，提出保护隐私的联合分布式深度学习方案，并且提供差分隐私保障。每个参与训练人员将本地的一小部分梯度参数引入噪声后上传到中心参数服务器端，每次更新本地参数时从服务器下载部分最新的梯度参数进行更新，使参与者在不公开自己数据集的情况下，从其他参与者中获益。2016 年，Liu 等[57]在此基础上使用 XMPP 作为参数服务器使之适用于移动分布式环境。2018 年，Le 等[58]针对文献[56]做出改进：如果参数服务器是好奇的，即使只上传一小部分梯度信息也会间接泄露用户隐私（如 4.1.2 中提到的通过逆向攻击恢复部分训练数据集信息）。作者提出在上传参数

时利用加法同态加密算法隐藏梯度信息。

除了加密技术和差分隐私技术，与 Shokri 和 Shmatikov 类似，2017 年，Google 推出联合学习[59-60]，Android 设备从 Google 服务器上下载模型，通过本地数据训练进行更新改进。与文献[56]不同的是，Google 并没有采用差分隐私技术，而是建议使用更安全的聚合协议安全多方计算（MPC），利用联合平均算法（federated averaging algorithm）[60-61]计算各个用户设备的更新，但 Google 服务器只有在多个用户参与时才能解密更新模型。Osia 等[62]提出一种用于隐私保护的移动分析混合深度学习架构。首次将孪生神经网络用于隐私保护的深度学习中，利用本地移动资源（如智能手机）提取敏感数据特征（预处理），再送入云端进行预测。既使用云端高效的资源，又避免直接暴露自身的敏感数据。

5　结束语

机器学习自身的脆弱性导致其安全威胁存在的必然性，并且容易泄露用户隐私，近年来，机器学习的安全与隐私问题引起了研究人员的广泛关注。机器学习被视为黑箱模型，其决策算法具有不可解释性，这给机器学习的安全防御和隐私保护带来一定的困难。当前，机器学习安全防御与隐私保护的研究仍处于起步阶段，还有许多亟待解决的问题。

1) 建立完善的评估机制。目前还没有统一的安全评估标准，对于隐私泄露没有统一的衡量标准，建立完善的评估机制、规范隐私保护规则是保障机器学习安全与隐私的重要一环。

2) 寻求有效的对抗训练方法。对抗训练的非适应性使在对抗训练中必须引入足够多样的对抗样本才能有效防御未知的对抗攻击，这是对抗训练的难点，有待解决。

3) 高效的加密方法保护用户隐私。最直接有效保护隐私的方法是使用加密技术，但目前同态加密技术运算开销过大，并不能直接计算机器学习中的一些非多项式运算。通常保护用户隐私是以牺牲目标模型的精度为代价。因此，研究高效的加密方法保护用户隐私是一个重要研究问题。

参考文献：

[1] GHORBEL A, GHORBEL M, JMAIEL M. Privacy in cloud computing environments: a survey and research challenges[J]. Journal of Supercomputing, 2017, 73(6):2763-2800.

[2] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[3] BARRENO M, NELSON B, SEARS R, et al. Can machine learning be secure?[C]//ACM Symposium on Information, Computer and Communications Security. 2006:16-25.

[4] KEARNS M, LI M. Learning in the presence of malicious errors[J]. SIAM Journal on Computing, 1993, 22(4): 807-837.

[5] BIGGIO B, NELSON B, LASKOV P. Support vector machines under adversarial label noise[J]. Journal of Machine Learning Research, 2011, 20(3):97-112.

[6] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[C]//International Coference on International Conference on Machine Learning. 2012: 1467-1474.

[7] MEI S, ZHU X. Using machine teaching to identify optimal training-set attacks on machine learners[C]//AAAI. 2015: 2871-2877.

[8] BIGGIO B, DIDACI L, FUMERA G, et al. Poisoning attacks to compromise face templates[C]//International Conference on Biometrics. 2013: 1-7.

[9] KLOFT M, LASKOV P. Security analysis of online anomaly detection[J]. Journal of Machine Learning Research, 2010, 13(1): 3681-3724.

[10] C. SZEGEDY, W. ZAREMBA, I. SUTSKEVER, et al. Intriguing properties of neural networks[C]//2014 International Conference on Learning Representations. Computational and Biological Learning Society. 2014.

[11] PAPERNOT N, MC D P, SINHA A, et al. Towards the science of security and privacy in machine learning[J]. arXiv preprint arXiv: 1611.03814, 2016.

[12] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. 2015.

[13] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[J]. arXiv preprint arXiv:1611.01236, 2017.

[14] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[J]. arXiv preprint arXiv:1710.06081, 2017.

[15] MIYATO T, MAEDA S, KOYAMA M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. arXiv preprint 1704.03976, 2017.

[16] MOOSAVI-DEZFOOLI S, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2574-2582.

[17] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//IEEE European Symposium on Security and Privacy. 2016:372-387.

[18] SU J, VARGAS D V, KOUICHI S. One pixel attack for fooling deep neural networks[J]. arXiv preprint arXiv:1710.08864, 2017.

[19] LOWD D, MEEK C. Adversarial learning[C]//The eleventh ACM SIGKDD International Conference on Knowledge Discovery in

Data Mining. 2005: 641-647.

[20] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[21] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//2017 ACM on Asia Conf on Computer and Communications Security. 2017:506-519.

[22] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples[J]. arXiv preprint arXiv: 1605.07277, 2016.

[23] GU S X, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[J]. arXiv preprint arXiv:1412.5068, 2014.

[24] LYU C, HUANG K, LIANG H N. A unified gradient regularization family for adversarial examples[C]//IEEE International Conference on Data Mining. 2016:301-309.

[25] ZHAO Q Y, GRIFFIN L D. Suppressing the unusual: towards robust cnns using symmetric activation functions[J]. arXiv preprint arXiv:1603.05145, 2016.

[26] ROZSA A, GUNTHER M, BOULT T E. Towards robust deep neural networks with BANG[J]. arXiv preprint arXiv:1612.00138, 2016.

[27] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. Computational and Biological Learning Society, 2015.

[28] HUANG R, XU B, SCHUURMANS D, et al. Learning with a strong adversary[J]. arXiv preprint arXiv:1511.03034, 2015.

[29] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. ensemble adversarial training: attacks and defenses[J]. arXiv preprint arXiv: 1705.07204, 2017.

[30] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]// IEEE Symp on Security and Privacy. 2016:582-597.

[31] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv: 1503.02531, 2015.

[32] PAPERNOT N, MCDANIEL P. Extending defensive distillation[J]. arXiv preprint arXiv:1705.05264, 2017.

[33] BULÒ S R, BIGGIO B, PILLAI I, et al. Randomized prediction games for adversarial machine learning[J]. IEEE transactions on neural networks and learning systems, 2017, 28(11): 2466-2478.

[34] HARDT M, MEGIDDO N, PAPADIMITRIOU C, et al. Strategic classification[C]//2016 ACM conference on innovations in theoretical computer science. 2016: 111-122.

[35] BRÜCKNER M, KANZOW C, SCHEFFER T. Static prediction games for adversarial learning problems[J]. Journal of Machine Learning Research, 2012, 13(Sep): 2617-2654.

[36] METZEN J H, GENEWEIN T, FISCHER V, et al. On detecting adversarial perturbations[J]. arXiv preprint arXiv: 1702.04267, 2017.

[37] LU JIAJUN, ISSARANON T, FORSYTH D. SAFETYNET: Detecting and rejecting adversarial examples robustly[J]. arXiv preprint arXiv: 1704.00103, 2017.

[38] HITAJ B, ATENIESE G, PEREZ-CRUZ F. Deep models under the GAN: information leakage from collaborative deep learning[C]// ACM Sigsac Conference. 2017: 603-618.

[39] FREDRIKSON M, LANTZ E, JHA S, et al. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing[C]//The 23rd Usenix Security Symposium. 2014: 17-32.

[40] FREDRIKSON M, JHA S, RISTENPART T. Model inversion attacks that exploit confidence information and basic countermeasures[C]//The 22nd ACM Sigsac Conference on Computer and Communications Security. 2015:1322-1333.

[41] ATENIESE G, MANCINI L V, SPOGNARDI A, et al. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers[J]. International Journal of Security and Networks, 2015, 10(3):137-150.

[42] SHOKRI R, STRONATI M, SONG C, et al. Membership inference attacks against machine learning models[J]. arXiv preprint arXiv: 1610.05820, 2016.

[43] TRAMER F, ZHANG F, JUELS A, et al. Stealing machine learning models via prediction apis[J]. arXiv preprint arXiv:1609.02943, 2016.

[44] GENTRY, CRAIG. Fully homomorphic encryption using ideal lattices[J]. Stoc, 2009, 9(4):169-178.

[45] DOWLIN N, RAN G B, LAINE K, et al. CryptoNets: applying neural networks to encrypted data with high throughput and accuracy[C]//Radio and Wireless Symposium. 2016:76-78.

[46] HESAMIFARD E, TAKABI H, GHASEMI M, et al. Privacy-preserving machine learning in cloud[C]//The 2017 on Cloud Computing Security Workshop. 2017: 39-43.

[47] BARYALAI M, JANG-JACCARD J, LIU D. Towards privacy-preserving classification in neural networks[C]//IEEE Privacy, Security and Trust. 2017: 392-399.

[48] XIE P, BILENKO M, FINLEY T, et al. Crypto-nets: neural networks over encrypted data[J]. Computer Science, 2014.

[49] STONE M H. The generalized weierstrass approximation theorem[J]. Mathematics Magazine, 1948, 21(4): 167-184.

[50] ZHANG Q, YANG L, CHEN Z. Privacy preserving deep computation model on cloud for big data feature learning[J]. IEEE Transactions on Computers, 2016, 65(5): 1351-1362.

[51] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//The Third conference on Theory of Cryptography. 2006: 265-284.

[52] ABADI M, CHU A, GOODFELLOW I, et al. Deep learning with differential privacy[C]//2016 ACM Sigsac Conference on Computer and Communications Security. 2016: 308-318.

[53] PAPERNOT N, ABADI M, ERLINGSSON U, et al. Semi- supervised knowledge transfer for deep learning from private training data[J]. arXiv preprint arXiv:1610.05755, 2016.

[54] BEAULIEUJONES B K, WU Z S, WILLIAMS C J, et al. Privacy-preserving generative deep neural networks support clinical data sharing[J]. bioRxiv, 2017.

[55] 郭鹏, 钟尚平, 陈开志, 等. 差分隐私 GAN 梯度裁剪阈值的自适应选取方法[J]. 网络与信息安全学报, 2018, 4(5):10-20.
GUO P, ZHONG S P, CHEN K Z，et al. Adaptive selection method

of differential privacy[J]. Chinese Journal of Network and Information Security, 2018, 4(5):10-20.

[56] SHOKRI R, SHMATIKOV V. Privacy-preserving deep learning[C]//The 22nd ACM SIGSAC Conference on Computer and Communications Security. 2015: 1310-1321.

[57] LIU M, JIANG H, CHEN J, et al. A collaborative privacy-preserving deep learning system in distributed mobile environment[C]//International Conference on Computational Science and Computational Intelligence. 2017: 192-197.

[58] LE T P, AONO Y, HAYASHI T, et al. Privacy-preserving deep learning via additively homomorphic encryption[J]. IEEE Transactions on Information Forensics & Security, 2018, 13(5):1333-1345.

[59] MCMAHAN B, RAMAGE D. Federated learning: collaborative machine learning without centralized training data[J]. Google Research Blog, 2017.

[60] BONAWITZ K, IVANOV V, KREUTER B, et al. Practical secure aggregation for privacy-preserving machine learning[C]//2017 ACM Sigsac Conference on Computer and Communications Security. 2017: 1175-1191.

[61] MCMAHAN H B, MOORE E, RAMAGE D, et al. Federated learning of deep networks using model averaging[J]. arXiv preprint arXiv:1502.01710v5, 2016.

[62] OSSIA S A, SHAMSABADI A S, TAHERI A, et al. A hybrid deep learning architecture for privacy-preserving mobile analytics[J]. arXiv preprint arXiv:1703.02952, 2017.

[作者简介]

宋蕾（1989-），女，黑龙江牡丹江人，哈尔滨工程大学博士生，主要研究方向为机器学习安全与隐私保护、云计算、网络安全。

马春光（1974-），男，黑龙江双城人，哈尔滨工程大学教授、博士生导师，主要研究方向为分布式密码算法与协议、云计算安全与隐私、格密码、机器学习安全与隐私保护。

段广晗（1994-），男，黑龙江海伦人，哈尔滨工程大学博士生，主要研究方向为深度学习、对抗样本、机器学习。