机器学习安全攻击与防御机制的相关研究

王文华
北京师范大学计算思维教育研究中心　北京　100037

摘　要：虽然机器学习的应用范围涉及人工智能的各个领域，但由于储存和数据传输安全问题和机器学习算法自身的缺陷，机器学习中存在着许多面对安全性和隐私权的威胁，以下根据攻击出现的情况和时间对机器学习中的安全性和隐私权威胁加以了划分，剖析标签投毒、大数据投毒、白盒入侵、黑盒入侵等产生的起因和攻击方式，并阐述和剖析了现有的信息安全防御机制。

关键词：机器学习；安全防御；攻击

Research on machine learning security attack and defense mechanism

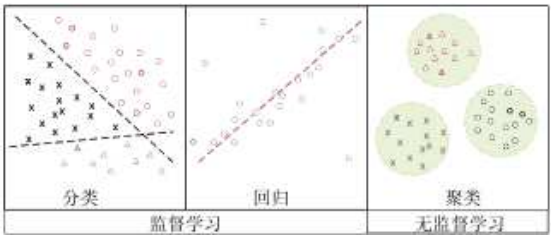
Wenhua Wang
Computing thinking education research center of Beijing Normal University, Beijing 100037

Abstract: Although the application scope of machine learning involves various fields of artificial intelligence, due to the security problems of storage and data transmission and the defects of machine learning algorithms, there are many threats to security and privacy in machine learning. The following divides the security and privacy threats in machine learning according to the situation and time of attacks, and analyzes label poisoning, big data poisoning, white box intrusion The causes and attack methods of black box intrusion, and expounds and analyzes the existing information security defense mechanisms.

Keywords: machine learning; Security defense; attack

一、引言

机器学习按照形式可分成监督教学、无监督教学和强化教学（如图一所示），督导教学的培训数据带有人为设置的标签，一般广泛应用于分级和预期任务，常用于统计学类型和回归模型进行分析，如垃圾邮件类型、房价预期等，在分级教学中，样本归属于二个或以上离散的类，建模的主要目标是将新的测试样本分配到这个类中，分类模型的建立主要是利用查找不同类中间的超平面，或利用支持向量机（SVM）、神经网络或逻辑回归分析等方式进行，回归分析建模指训练标签为连续值的培训测试样本生成的建模，利用拟合一种与培训测试样本的数据信息相距最近的建模（常常为一种曲线），能够达到对新的测试资料基于特征值的估计，当训练数据无法区分或人工方式设置类型成本费用太高时，只能采用机械学习处理模式识别中的各类问题称之为无监督学习，也叫做归纳性教学、无监督学习常见于聚类方法和训练数据降维，采用循环和递减算法来降低偏差，从而达到分类目标的，K-means是最常用的聚类方式之一强化学习建模，是根据奖励值在学习状态与培训动作间映射的建模，一般运用在学习策略选择中，如AlphaGO围棋机器人的策略增强学习模型按照最动态的奖励值调节其权重，等最新的状态出现时，再选择当前奖励值最高的学习策略。



图一　机器学习分类

作者简介：王文华，1993年07月05日，男，汉，甘肃，本科，专利工程师，研究的方向：人工智能（计算思维）领域专利开发。

二、机器学习面临的攻击

1.标签投毒

进攻者可以透过进行改变练习数据分析的标记讯息，将练习数据分析对接到出错的标记，模型掌握到了出错的相对关联性后，当接受新的试验数据分析时会背离通常判定，准确性下降，当袭击者已经掌握练习数据分析的权限讯息时，进行改变练习数据标签将是很轻易进行的攻击，Binggio等人叫在SVM分类练习机中，进行改变了百分之四十的练习数据标签，使得模式中仅有了百分之三十的正确性。

2.数据投毒

当练习信息流入模式之前，袭击者由于改变原来的练习信息或产生新的出错的信息发动污染射击，从而使得练习模式的精确度下降，而基于机器学习模型注入信息的特性，可将练习模式分成经常离线的练习模式和定期更新实时注入信息的练习模式，当模式为离线注入信息后，Biggio等人采用检测出错的固定公式的SVM学习算法，由于改变其练习信息，从而降低了练习精确度；Mei等人根据图损失练习（包括线性或逻辑回归或SVM练习和连续输入的模式，提出了更一般化的污染架构，认为如果模式适合，就可以找出最好的污染方法。

Kloft等对于网络或者要求经常改变的学习方式，建议采用基于模型中的常规方法获得的训练资料来影响新获得的训练资料，从而减少了系统的稳定性；他提出了一种新的网络学习模式，即根据网络上的实际情况，发现了不同于平均水平的培训资料，并对新的网络学习模式进行了修正，导致了系统的稳定性下降。

当威胁者无权读取预处理数据信息，却可能通过间接污染训练数据信息时，Perdisci等人在训练蠕虫标签产生器的数据信息流中添加了干扰，使其通过改变判别阈值大大降低了标签识别的准确度，也构成了数据投毒威胁。

3.白盒攻击

根据攻击者期望达到的目标，针对黑白盒子的对峙样品袭击可以包括错误类型、源/对象相反类型以及指向错误类型，这三种错误类型通常指二分类学习任务中，使结果出错的行为，通常出现在对非法软件、邮件或文档的鉴别教学任务中；源/目标所相应的错误分类通常指在多分类标记的错误分类任务中，使用优化算法产生的原始数据得到最小修改的抵抗样品，可以使其透过机器学习模型获得进攻者所指定的错误类型，通常出现在图像识别的任务中，而抵抗样品则可以获得次级可能性的错误标签类型；靶向错误分类攻击，通常指出攻击者所获得的对抗样本作为人类来说是无意义的，但是使用机器学习模型就可以进行攻击者所指定的类型，通常出现在图像识别或语言识别中，对无意义的图形或者语言特征进行了明确的区分或者标识。

4.黑盒攻击

目前的机器学习云服务平台多为使用者建立了试验用的查询端口，进攻者可以通过观察检测培训统计进入模式后所回复的结论实施进攻，但是黑箱攻击和白箱进攻有所不同，因为袭击者既没法判断所要干扰和错误划分到的目标类型，也没法得到训练统计和模式统计，所以没法基于背景知识设计并进行最优预测对付样本首先，面向机械学习任务的黑箱对付样本攻击是由Lowd等人叫提议，在垃圾邮件过滤器中以单词作为变数判断垃圾邮件类型的条件，经过反复问询后，为变数增加了标签，利用逆向工程使用成本函数来获取垃圾信息并进行过滤器的小改动后，完成了对付样品攻击，在回归式机器学习任务中，由Alfeld等人根据预测的回归模式，利用对模型的推理，并修正输入数值，以获得期望输出，基于袭击者的知识，如果袭击者的知识不同，当攻击对象知道了模式标记的几率后，攻击的效率只比白盒子攻击稍逊一筹，泽维尔等人通过PDF框架和Hidostl两种已有的PDF框架和Hidostl方法建立了一种新的机器学习模式，它相信只要将一些执行编码插入到pdf中，就可以实施攻击。而Rndic等人则认为：主动的防御方式仅对特定进攻方法才有效，如果袭击者运用其知识库就能够重构模型并训练数据，但是如果进攻策略稍有改变，则恶性的pdf识别效率大大降低，并提出了可以透过向pdF文档中添加被pdfrenderer可以省略的信息进行了攻击的例子和试验报告[1]。

5.成员推理攻击

攻击主体的攻击对象为攻击对象，攻击对象可以从攻击的角度判断攻击对象是否参与了模型的训练，攻击的攻击和统计的综合信息和有噪音的实际资料构成攻击模式，并且摧毁了机器学习的模型，保证了训练的隐私，Shokri等人则使用成员推理攻击者判定出某个疾病相关的数据个体是不是参加了模拟培训、并破坏了训练模型的私密性。

6.训练数据提取攻击

练习数据获取进攻者的主要目标是练习数据分析的条目数据信息，是指进攻者利用询问数据结果和现有知识来推断练习数据分析秘密的攻击，Fredrikson等人进一步认为：进攻者通过利用建模输出信息与某个特定属性

之间的关系，从而能够推断练习数据分析中的秘密讯息，文章利用建模信息输出与人口统计数据信息，从用药剂量估计模式中能够顺利回复练习数据分析中的基因讯息，Fredrikson等人进一步认为：利用建模反演进攻能够使进攻者获得大部分建模输入信息，而相对于建模进入，练习统计和分析的人口统计数据信息更具备秘密属性，在练习数据获取进攻中，进攻者通过利用大规模提问的结果获取训练模型的分类信息以及对各个类型输出的概率，并以此建立与模式类似的特性矢量，每种特性矢量值代表了某种类型的一般特性矢量值，当某种类型中仅有一名个人时，如果该个人隐私信息暴露于人脸识别模式中，攻击者就能够得到这个人的所有人脸消息[2]。

7.模型提取攻击

模型获得攻击的目的是通过对机器学习模型进行训练，然后通过询问界面获得原始模型的分析与试验结果，从而重建一个与模型相似的模型，Vorobeychik等人通过现有的模型和询问能力，对模型进行建模，得到了与模型类型相同但模型未知的模型，并运用这种数据属性构建一种全新的训练模型，以便进而实现训练数据获取进攻，Tramer等人还认为：通过问询接口，光靠观察原模型推测过程中的输入输出信息，对进攻者来说也能够获取原模型信息，甚至建立一种相似的训练辅助模型用以实施进一步的攻击[3]。

三、安全防御机制及分析

（一）模型正则化

所谓模式规范化，就是通过使用规则化项使模型的资料与训练方式规范化，以改善模型的推广效果，Barreno等人指出，尽管DAE（DAE）可以有效地去除某些抗干扰的干扰，但是较少的改进后的新的对抗性样品仍然可以有效地抵抗自解码和深层神经网络。在损失函数中增加了平滑度惩罚，即正则化项，旨在最小化经验风险的同时，尽量减少微小改变对模型输出的影响，并增加建模对样本的鲁棒性。

（二）对抗训练机制

面对样本进攻的形成，主要是因为机器学习模型提供的数据维度太高而训练模型不够线性所造成的结果，即是训练模型泛化能力不够，所以根本无法完全掌握到练习数据分析与目标标签的映射关联，在特定的培训技术背景认知下，才能够利用增加一些干扰生成对抗样本来突破训练模型的决策界限，从而实现了抵御进攻的目的，为预防面对样品进攻，除提高了培训数据质量以外的研究人员在练习数据分析与建模改进二方面，先后形成了抵抗训练与防御精馏等的安全模式。

对立练习，即是通过反相模型生成一个具有完整标记的反面样本，和一个法定样本结合在一起，可以增强一个强大的防御系统，例如PinTo等，他们通过一种类似于“互相攻击”的方式，尝试着去捕捉一个对象，而另外一个则尝试破坏它的均衡，这就是Kurakin等所说的：利用转移，可将从小的数据库系统的对立样品练习推广到大数据库的对立样品练习，Goodfellow等人使用对立练习，使在MNIST数据集上的错误识别率由八十九，百分之四减至十七，百分之九，Huang等人使用惩罚错分类的对立样品练习，提高了模式的顽健性，tRamer等人提供了联合对抗训练（ensEmbleadversarialtraining），以提高抵抗样本多样性，但同时也指出：在对抗训练中导入有未知攻击目标的抵抗样品是不实际的，而对抗训练的非适应性也造成了对抗训练的局限[4]。

从抵抗样本的存在问题被提出，研究依据建模特性给出了L-BFGS、FGSM、DeepFoo、Carlini-Wagner等人使用的各种抵抗样品形成计算，Xu等提出，利用基因演算法可以进一步制造出新的抗药性样本，而对抗性培训的目的是将新的抗药性样本与合适的标记相结合，由此可以了解模型特征与合适的对应关系，由此发现更多的抗样本生成的运算，由此生成和维持充足数目的新的抗性样本，从而能够更有效抵御反抗样品的威胁，从而改善了建模特性。

（三）防御精馏

精馆是利用一种模式的输出来练习另一模式的机器学习算法，并在进一步提高练习精确度的前提下缩小模式的方法，而防守精馏则是由Papernot等人在训练精馏方法的基础上，利用两种相同模式间的练习，实现了梯度掩码，并以此改善模式面对对抗样本的鲁棒性的方法，在后来的科学研究中，Papernot等人认为，针对黑盒攻击防守精馏方法具有缺陷，并给出了可推广的防守精馏方法，试验结果表明：通过防守精馏方法能够生成比输出表面更加光滑的、对扰动较不灵敏的模式从而改善了模式的顽健性，并可以使对样本入侵的几率由九十五型%减至低于百分之零点五，但很快，Carlini等又指明了通过防止精馏出现问题，并提供了大规模破坏性防御和馏安全性的威胁。

（四）输入正则化

如果你的模拟能力很强，但是训练次数太少，很可能会出现错误，因为模拟训练很可能会出现错误，所以训练的次数越多越好。在有充足的培训模型的情况下，

攻击者可以通过干扰训练的数据来降低训练的质量，从而导致错误的产生，降低仿真的准确性。所以为了提高训练模拟的性能，就一定要提高训练数据的质量，高质量的训练数据具有合适的特征空间和数据分布，并且具有适当的数量，对训练数据质量的正则化也可理解为在确保了训练数据存储安全性的情形下，通过增加训练数据的质量增加训练数据的质量也叫做数据集增强，也就是通过利用特征空间提取改变训练数据集合的特征空间和数据分布信息，并利用注入噪声实现数据扩展，进而产生更新的训练样本，从而产生具有更大容量或者无限容量的增强训练数据集，进而提高模型的一般化能力机器学习家认为，更多的练习数量能够减少模型犯错量，使模型更具有一般化能力，因此，扩展练习数据集是增强建模性能的主要手段，而袭击者通过改变练习数量。例如，在训练数据的垃圾信息中添加一个词语，或从检查测试数据的垃圾信息中排除这个词语的使用，就可以避免检查或测试训练数据，因此，对垃圾信息攻击者通常会添加积极地或者具有隐含意味的词汇来避免垃圾邮件过滤地分布，就能够实现投毒攻击或者对抗样本攻击的目的，这源于模型对未知数据不存在鲁棒对于防御投毒攻击[5]。

四、结语

多数集中式教学的防守机理都构建在寻找培训不在期望的范围内的样本上，以增强训练模型面对不明数据信息的对抗力量，如Rubinstein等人就认为：规范了培训的数据分布空间结构，以减低投毒进攻者的对模型进入期望，Biggio等人也通过正则化输入空间以减低因进攻者更改培训标签造成的逃逸威胁，而分布式教学的防守机理则构建在寻找参与者所训练出的不在于期内的样本上。

参考文献：

[1]李欣姣，吴国伟，姚琳，等.机器学习安全攻击与防御机制研究进展和未来挑战[J].软件学报，2021，32（2）：18.

[2]孙歆，方芳，孙昌华.基于机器学习及特定攻击特征的多维度工业系统安全感知方法研究[J].电子设计工程，2019，27（17）：6.

[3]王前，王磊，谢寿生.DDoS攻击和防御机制分类研究[J].计算机应用研究，2006，23（10）：3.

[4]李盼，赵文涛，刘强，等.机器学习安全性问题及其防御技术研究综述[J].计算机科学与探索，2018，012（002）：171-184.

[5]Lin C，Honggang L，Yuanfei H，et al. Research on mechanism of resisting malicious code based on trusted computing基于可信计算的恶意代码防御机制研究[J].计算机应用研究，2008，25（12）：3713-3715.