敌对攻击环境下基于移动目标防御的算法稳健性增强方法

何康[1,2]，祝跃飞[1,2]，刘龙[1,2]，芦斌[1,2]，刘彬[1,2]
（1. 信息工程大学网络空间安全学院，河南 郑州 450001；
2. 数学工程与先进计算国家重点实验室，河南 郑州 450001）

摘 要：传统的机器学习模型工作在良性环境中，通常假设训练数据和测试数据是同分布的，但在恶意文档检测等领域该假设被打破。敌人通过修改测试样本对分类算法展开攻击，使精巧构造的恶意样本能够逃过机器学习算法的检测。为了提高机器学习算法的安全性，提出了基于移动目标防御技术的算法稳健性增强方法。实验证明，该方法通过在算法模型、特征选择、结果输出等阶段的动态变换，能够有效抵御攻击者对检测算法的逃逸攻击。

关键词：机器学习；算法稳健性；移动目标防御；动态变换
中图分类号：TP301.6
文献标识码：A
doi: 10.11959/j.issn.2096−109x.2020052

Improve the robustness of algorithm under adversarial environment by moving target defense

HE Kang[1,2], ZHU Yuefei[1,2], LIU Long[1,2], LU Bin[1,2], LIU Bin[1,2]
1. Cyberspace Security Institute, Information Engineering University, Zhengzhou 450001, China
2. State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China

Abstract: Traditional machine learning models works in peace environment, assuming that training data and test data share the same distribution. However, the hypothesis does not hold in areas like malicious document detection. The enemy attacks the classification algorithm by modifying the test samples so that the well-constructed malicious samples can escape the detection by machine learning models. To improve the security of machine learning algorithms, moving target defense (MTD) based method was proposed to enhance the robustness. Experimental results show that the proposed method could effectively resist the evasion attack to detection algorithm by dynamic transformation in the stages of algorithm model, feature selection and result output.

Key words: machine learning, algorithm robustness, moving target defense, dynamic transformation

收稿日期：2019−11−26；**修回日期**：2020−02−04
通信作者：祝跃飞，yfzhu17@sina.com
基金项目：国家重点研发计划基金（2016YFB0801505）； 国家重点研发计划前沿科技创新专项基金（2019QY1305）
Foundation Items: The National Key R&D Program of China (2016YFB0801505), Cutting-edge Science and Technology Innovation Project of the Key R&D Program of China (2019QY1305)

论文引用格式：何康, 祝跃飞, 刘龙, 等. 敌对攻击环境下基于移动目标防御的算法稳健性增强方法[J]. 网络与信息安全学报, 2020, 6(4): 67-76.
HE K, ZHU Y F, LIU L, et al. Improve the robustness of algorithm under adversarial environment by moving target defense[J]. Chinese Journal of Network and Information Security, 2020, 6(4): 67-76.

1 引言

近年来，机器学习算法在数据挖掘、图像处理、自然语言处理和网络空间安全等方面得到了广泛应用，分类准确率上取得了显著提升。但面对以入侵检测、恶意软件检测为代表的网络安全领域[1-4]的分类任务时，算法在提高算法分类准确率的同时，还需应对敌人对算法本身的攻击。传统的机器学习任务通常假设训练数据集和测试数据集的特征服从基本一致的概率分布，进而取得较高的准确率。基于流形学中关于低概率区域的观点认为，由于训练样本的有限性，机器学习的训练过程只学习了总体样本所在的概率空间的局部区域，而对抗样本是从总体样本所在的概率空间的某一子空间抽样得到，其超出了分类器所能学习的概率分布所在的支集[5]，因而攻击者会通过精巧修改输入样本来打破该假设并误导分类算法输出期望的错误判断。在图像处理[6]、语音识别[7]等领域，已有对输入样本数据的小幅改动能够逃逸分类算法检测的研究，在网络安全领域，在检测模型算法上的对抗也日趋激烈[8-9]。

常规的机器学习防御方法通过添加对抗样本到训练集中重训练模型优化分类性能，但给定足够长的时间，攻击者总能构造出不在训练数据分布范围内的对抗样本。Nicolas 等[10]提出训练过程中的防御性蒸馏增加了输出类别之间的相似性信息，提高了算法的稳健性，但 Carlini 等[11]的研究表明，攻击方式稍作修改仍能有效降低算法的准确率。

基于此，本文基于移动目标防御（MTD，moving target defense）思想[12]提出一种算法稳健性增强的方法。传统的防御系统只是对防御方的系统进行加固，但无法保证加固后的系统信息不暴露给攻击者。攻击者经过不断地试探、分析、研究，能不断获取目标信息，攻击者有充分的时间和资源挖掘系统的弱点，而防御者对攻击者的信息知之甚少，无法进一步提高防御水平，攻击者对防御者具有天然的不对称优势。为了打破这种不对称优势，增加防御者的动态性和随机性，MTD 技术应运而生。其目标是通过构建一个动态变化、冗余异构、具有不确定性的防御系统，使攻击者的目标一直处于动态变化过程中，大大增加了攻击者的攻击代价。MTD 技术在操作系统地址空间和指令集[13]、网络地址和端口[14]等方面的应用已显著提高了网络系统的安全性，其防御思想可引入算法系统的防御中。MTDeep[15]将 MTD 思想应用到神经网络算法中，张红旗团队[16]使用主从博弈对其效能进行了分析。Roy[17]、李亚龙[18]等将该思想进一步拓展到非理性条件和非完全信息情况下的攻防博弈分析。攻防双方采取不同概率分布策略时的随机预测博弈模型，以提高攻击开销。

本文首先对 MTD 在算法上的应用进行形式化，做出对稳健性增益的定性分析；然后将 MTD 应用到算法设计的多个阶段，并给出了具体的实施方法；最后在恶意 PDF 分类领域对算法稳健性的增益进行检验。基于 MTD 的动态算法系统机制不是代替其他算法稳健性增强的方法，而是从额外的维度为现有的算法稳健性设计方法提供强有力的补充。

2 算法分析

为对采用 MTD 技术算法的防御能力进行分析，首先对攻击者的能力进行假设。网络安全领域机器学习算法的典型应用包括对软件或者文档的恶意性进行分类，如 VirusTotal、VirusShare 等网站。用户仅可提交样本并获取查询结果，对其中涉及算法的训练集、分类器形式、参数等知识了解甚少，因此假设攻击方式为黑盒攻击。具体过程如下。

数据集 $D=(X,Y)$，其中，X 为特征向量，Y 为样本的类标签向量。D 经过机器学习训练得到模型 M_1 的过程，表示为 $M_1=\text{train}(X,Y)$，输入数据 X 经过 M_1 分类的结果表示为 $Y'=\text{predict}(M_1,X)$。攻击者使用自有数据集 $D'=(X',Y')$ 经过 M_1 分类得到结果 $Y''=\text{predict}(M_1,X')$。由于攻击者对目标模型的了解较少，攻击者使用查询数据训练一个新的模型 $M_2=\text{train}(X',Y'')$ 作为替代目标展开攻击。这样做，一方面因为直接对线上分类器进行多次查询容易引起防御方的警觉，另一方面因为样本的可

迁移性。样本的可迁移性是指在训练集相同的情况下，被模型 M_1 错误分类的样本可能被模型 M_2 错误分类。这样攻击者可以充分利用模型 M_2 的知识展开攻击，并将成功取得的对抗样本应用到对模型 M_1 的攻击上。尽管如此，由于不同模型参数以及分割面特性的差异，样本在不同模型上的分类结果也不尽相同。MTD 在算法设计上的目标就是对抗模型的可迁移性。对样本的可迁移性定义如下。

定义 1　单个样本 x 在分类模型集 M 上的可迁移性：对多个分类模型组成的集合 $M=\{M_1, M_2, \cdots, M_n\}$，样本 x 在每个模型下分类的类别数为 m，则单个样本的可迁移性定义如式(1)所示。

$$P(j)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}_j(\mathrm{predict}(M_i,x)) \tag{1}$$

其中，$P(j)$ 表示样本 x 被所有模型分类为类别 j 的比例，$I_j(w)$ 为示性函数，定义如式(2)所示。

$$I_j(w)=\begin{cases}1, j=w\\0, j\neq w\end{cases} \tag{2}$$

$$\mathrm{trans}(M,x)=\sum_{j=1}^{m}P(j)\cdot\log(P(j)) \tag{3}$$

该定义的形式借鉴了熵的概念，$\mathrm{trans}(M,x)\leqslant 0$。如果 M 中所有的模型对样本 x 的分类均相同，那么说明 x 在模型之间具有较高的可迁移性；如果各模型对样本 x 分类的结果差异较大，说明 x 在模型之间的迁移性较低。样本的可迁移性反映了样本在不同模型上分类结果的差异性。

定义 2　样本集 D 在分类模型 M 上的平均可迁移性如式(4)所示。

$$\mathrm{aver\_trans}(M,D)=\frac{1}{|D|}\sum_{x\in D}\mathrm{trans}(M,x) \tag{4}$$

其中，$|D|$ 代表样本集 D 的大小，平均可迁移性反映整个样本集在模型集上分类结果差异的平均水平。

定义 3　给定模型可以正确分类的样本 x、模型 M、误差范围 ε，如果

$$\begin{aligned}&\exists x'\text{满足约束条件}|x-x'|<\varepsilon\cup\\&\mathrm{predict}(M_i,x)\neq\mathrm{predict}(M_i,x')\end{aligned} \tag{5}$$

则称 x' 为 x 的对抗样本。

攻击者的目标是在保留恶意功能的前提下产生能够欺骗分类算法的对抗样本。对抗样本产生的一个原因是训练集 D_1 是样本空间 Ω 的一个子集，训练阶段所学到的算法仅能够对训练集所在的子空间有效。而对抗样本构成的数据集 D_2 超出了 D_1 所在的概率分布所在的支撑集，因而算法无法对对抗样本做出有意义的判断。本文进一步将测试数据集划分为核心数据集 C 和边缘数据集 E。C 中的样本分布在算法能够做出有效分类的样本子空间内，相对地，边缘数据集 E 的分布超出了训练集能够表示的概率空间，其分类结果的参考意义较小。对抗样本一般出现在边缘数据集 E 中，可以进一步定义如下。

定义 4　模型集合 M 上的边缘数据集 E 与核心数据集 C 分别如式(6)、式(7)所示。

$$E(L)=\{x\in D\,|\,\mathrm{aver\_trans}(M,D)<L\} \tag{6}$$

$$C(L)=\{x\in D\,|\,\mathrm{aver\_trans}(M,D)\geqslant L\} \tag{7}$$

其中，L 为给定的区分 E 和 C 的界限。

根据上述定义，可以得出以下性质。

性质 1　对样本集 D、模型集 M，当 $L=0$ 时，核心数据集 $C(L)$ 中的样本被 M 中的模型分类的结果均相同。

性质 2　若对抗样本集 D 被模型集 M 划分为 $C(0)$ 和 $E(0)$，则动态切换 M 中的模型不会对 $C(0)$ 中的样本产生不同的分类结果。

性质 3　假设对模型 P 生成的对抗样本集 D，即 D 中的任意样本 x 在模型 P 下均产生错误分类，特别地，将恶意样本分类为良性样本的数据。对模型集 $M=\{M_1, M_2, \cdots, M_n\}$，令 $C_i=\{x\in D\,|\,\mathrm{predict}(M_i,x)=\mathrm{predict}(P,x)\}$，则 D 的核心数据集 $C(0)$ 的含义是可以骗过 M 中所有模型的对抗样本集，表示为式(8)。

$$C(0)=\bigcap_{i=1}^{n}C_i\cap D \tag{8}$$

性质 4　如果假设每个模型对对抗样本集 D 中的判定是独立的，则 D 中的数据被 M 中所有模型划分为相同分类的概率 P 如式(9)所示。

$$P=\frac{|C(0)|}{|D|}=\prod_{i=1}^{n}\frac{|C_i|}{|D|}=\prod_{i=1}^{n}P_i \tag{9}$$

P_i 表示 D 被模型 M_i 分类为目标类别的概率。

从上述内容可以看出，随着模型集 M 规模扩大，$C(0)$ 的大小是单调递减的，而且 M 中模型的差异性增大，$C(0)$ 的降幅也会增加。在假定判别模型相互独立的情形下，随着 n 的增加，被所有模型分类为指定结果的概率也迅速下降。虽然实际应用中模型的独立性假设难以完全满足，但能对 MTD 在算法设计上的有效性做出定性分析，即随着算法设计多样性的增加和算法各模块的动态变换，算法整体的稳健性不断增强。算法的 MTD 就是通过训练异构模型构成的模型集 M，并在预测分类结果时进行动态变换，实现对 D-$C(0)$ 部分对抗样本的有效预警。

3 动态算法系统的多样性设计

面对未知的固定不变的目标算法，攻击者有充足的时间和资源寻找可能存在的漏洞，而防御者对攻击者采取的行动一无所知，攻击者对防御者有天然的不对称优势。MTD 技术有 3 个典型特点：多样性、动态性、异构性。移动是在异构的配置中进行动态切换，减少算法的脆弱性在攻击者视野中的暴露时间，使攻击者的攻击速度落后于配置切换速度，达到有效增加攻击者的难度和代价的目的。在动态算法系统中，攻击者的目标就是执行分类任务的分类器，即机器学习算法模型本身。目标的集成以及动态切换构成了整体的动态算法系统。在本文中，算法的含义界定为根据训练数据生成的将输入向量映射到特定类别或属于各类别概率的执行分类任务的具体机器学习模型，不同的配置包括算法种类、使用的特征、参数和超参数等，是动态算法系统执行配置变换的基本单元。其中，多样性是实现 MTD 的基础，异构性是 MTD 能够有效工作的前提，如果算法模型之间没有差异性，则移动并不能带来算法稳健性上的增强。由于算法模型具有可迁移性的特点，如何在保证分类准确率的同时对算法进行多样化设计也是一个需要研究的重要问题。本文对这几个方面进行设计。

3.1 算法选择的多样性

对同一个机器学习任务，可以使用不同的算法模型进行实现，但不同算法在数据的分类上展现出不同的优缺点。如 K近邻算法可解释性强，但对训练数据集的依赖性较强，容易受到敏感值的影响；基于条件概率的贝叶斯算法对不同维度数据的独立性有较高的要求；支持向量机算法的分类面主要受支持向量的影响，对偏离数据重心的样本的判断，置信度不高，且不同核函数的选择对分类结果的影响也较大；神经网络算法准确率较高，但参数较多，容易产生过拟合等。不同的分类算法在数据集上产生的分类面有较大差异，因而对核心数据集和边缘数据集的划分也不同。算法模型本身的多样性有利于提高算法系统的防御能力。

3.2 特征选择的多样性

特征选择是从对象的大量描述维度中选取一部分用于机器学习训练的过程。该过程一方面能够去除重复冗余的特征来减少特征之间的相关性，另一方面通过降低维度来防止维度灾难引起的过拟合等问题，增强机器学习模型的泛化能力。不同的模型描述对象的角度不同，对数据的特征选择也有不同的侧重，因而机器学习算法使用各自特征集训练的模型在对输入的判断上也有所差别。在特征选择上多样性设计可以从不同的角度对数据集进行判别，降低由于攻击者对特征知识的判断生成对抗样本的能力。

3.3 算法输出的多样性

在对机器学习模型的攻击中，攻击者通过查询黑盒机器学习模型的输出构造新的训练集并生成训练模型。攻击者通常使用梯度下降等方法对模型中的参数进行优化，使模型的输出接近指定的类标签或者隶属度概率。一个直观的想法是，通过对模型的输出添加随机扰动，以干扰基于梯度下降的参数优化的方向，增加攻击者训练的替代模型与目标模型的差异性。如模型 M 对样本 x_1 分类为恶意的概率为 0.9，但加上 0.1 的扰动并不影响分类的最终结果。对处于分类边界的样本 x_2，扰动的添加虽然有可能改变样本的分类，但隶属度处于边界位置的样本本身的标签就具有不确定性。该方法通过扰动输出的方法改变处在边缘数据集中的分类输出结果和核心数据集的置信度，干扰了基于梯度优化算法的替代模型的训练和对抗样本的生成。

算法输出的多样性会对模型分类结果的

准确率造成影响，但影响较小。对训练良好的机器学习模型来说，非对抗样本的隶属度概率绝大多数靠近 0 或者 1，分布在分类界限 0.5 附近的样本数量很少。而对抗样本完全模拟正常样本的难度较大，有较大概率分布在分界线附近，分类概率输出的多样性会对对抗样本造成较大影响，进而给攻击者的替代模型造成较大困难。

以上列举了算法多样性设计的几种方式，在实际训练过程中可根据需求对算法模型其他方式的多样化，如使用不同的训练数据集和训练权重产生具有不同偏好的机器学习模型等。然而，算法的多样性不等同于集成学习算法。集成学习算法通过结合不同分类器的预测结果来产生最终的结果，但这些分类器本身是固定不变的，攻击者仍能通过长时间的信息反馈生成集成学习算法的对抗样本。Kantchelian 等[19]提出的基于混合整数线性规划的方法能对随机森林算法展开有效攻击并生成对抗样本。而基于 MTD 的算法系统可以动态切换使用机器学习模型，当攻击者针对当前模型生成对抗样本时，所面临的算法系统已经切换到新的机器学习模型，打破了攻击者对防御者的时间不对称优势，有效增强了算法面对敌对攻击的稳健性。

相比传统单一的检测算法，基于移动目标防御的动态算法系统地增加了部分开销，其成本有以下 4 个方面。① 设计成本：防御者需要根据任务要求设计冗余异构算法，在保证准确率的同时压缩对抗样本的概率分布空间。② 训练成本：使用数据集和算法对模型参数进行训练，单个模型的训练时间从分钟级到天级不等，算法池的规模为 N，则算法系统的训练成本为 $O(N)$，但该过程可以通过并行算法将训练成本降低到 $O(1)$。在实际部署中，算法系统可以在运行时增量训练，减少占用的时间成本，而模型的空间开销为 $O(N)$。③ 管理成本：动态算法系统需要对算法的工作状态进行管理和变换，需要占用一定的内存空间。④ 决策成本：相比单个算法直接获取检测结果，动态算法系统的决策模块需要集成多个算法的输出给出综合判断并为算法训练提供反馈。总体来说，动态算法系统对时间和空间占用控制在 $O(N)$，在时间和空间上不会造成太大的开销。

4　动态算法系统的流程设计

为了清晰地描述算法系统的动态变换过程，首先对相关概念做出说明。在算法设计流程中，在进行多样化设计的每个环节中，定义配置。

定义 5　配置 d=(name, value)，name 为配置参数的标识符，value 为该配置的取值。对每个配置 d，value 的定义域为 V。

定义 6　配置集 $D=\{d_1,d_2,\cdots,d_n\}$。算法设计中可多样化的位置的配置 d_i 构成配置集 D，其描述的是算法的整体多样化能力。整个配置集的定义域为各配置参数定义域的笛卡尔积，即 $V=V_1\times V_2\times\cdots\times V_n$，MTD 可变换的配置空间大小为 $\Omega=|V_1|\times|V_2|\times\cdots\times|V_n|$。

定义 7　算法模型 $m=(s_1,s_2,\cdots,s_n)$，为 D 中各配置取特定值构成的算法模型，是算法实现的具体描述。每个 m 对应一个检测模型。

定义 8　算法模型集 $M=\{m_1,m_2,\cdots,m_l\}$，为多个算法模型构成的集合，防御者根据 M 中模型对样本预测的综合结果给出判断和决策。

定义 9　动作序列 $A=\{a_1,a_2,\cdots,a_p\}$，为对算法模型集的操作序列，a_i 为对某个算法模型的操作。对算法系统来说，操作可分为两类：删除和添加模型。当认为模型 m_j 的输出与绝大多数模型的输出一致时，考虑到模型冗余性，可以考虑将该模型删除。当各模型的预测结果差异较大时，对样本的预测结果有较低的置信度，考虑增加新的模型来提供额外的信息。此外，对模型的操作还需依据算法系统的模型切换策略。

定义 10　模型约束集 $C=\{c_1,c_2,\cdots,c_k\}$，为确保模型分类的准确率，需要对算法模型集的选择保持一定的约束，即保留一批准确率较高、稳健性较强的算法模型作为基本模型；同时为了提高分类的稳健性，需要对算法模型进行多样化。

定义 11　策略集 $P=\{p_1,p_2,\cdots,p_z\}$，定义算法系统中 M 的动态变换策略，策略集 P 涉及两个方面：变换时机和变换目标。变换本身则通过动作序列 A 来完成。

定义 12　随机化方法集 $R=\{r_1,r_2,\cdots,r_n\}$，定义了算法系统中 D 变换所使用的随机策略，包括均匀分布、指数分布以及指定的马尔可夫链概率

转移模型等，在算法的实际应用中可根据需求设计不同的随机化方法。

变换时机方面，本文提出 3 种变换策略。① 定时变换：算法系统经过指定数量的样本检测或指定的运行周期后强制变换其中的一批算法，增大攻击者通过长时间对算法系统模型的查询操作生成替代模型的开销。② 触发变换：当输入算法系统的样本经过 M 的检测后出现较多不一致的结果时，该样本被判定为边缘数据集，可能需要额外的模型提供信息做进一步的检测。③ 随机变换：为了降低算法系统变换规则的规律性，在上述两种策略之外，使用 R 中的随机化因子触发算法系统中模型的变换。

变换目标方面，所选取的目标算法模型首先要满足策略集 P 约束。可以采用的策略如下。① 随机选取：在候选集中随机选取替代机器学习模型。② 最大化 D 的差异选取：从候选的机器学习模型中选取 x，使其配置集 D_x 与被替代模型的配置集 D_0 的 L_x（包括 L_0、L_1、L_∞ 等）距离最大化。在实际应用中，算法设计者可根据实际需求设计不同的变换时机和变换目标策略。

变换方式方面，可根据 D_x 与 D_0 的差异自动生成动作序列 A，对 M 中的模型做添加删除操作。

根据上述对动态算法系统的描述，设计的基于 MTD 技术的动态算法系统流程如图 1 所示。首先对算法进行多样化设计，并和训练数据集在分类准确率较高的前提下生成异构的机器学习模型池，供算法决策模块选择算法模型集 M。实际环境中的样本首先通过特征提取生成特征向量，然后经过 M 的检测输出 n 个检测结果，综合判别模块根据投票等决策机制输出对该样本的最终判别结果。M 的输出也会影响算法决策模块，检测前需运行算法决策模块决定是否对现有算法模型集进行变换。

算法决策模块的工作流程如图 2 所示。首次运行时，先从模型池中随机选取满足约束集 C 的初始模型集 M 作为输出。后续运行时，根据前述的变换策略决定是否对现有的 M 进行变换。当决定变换时，根据目标选取策略集 P 选择满足约束集 C 的替换模型并生成动作序列 A。执行变换后得到更新后的算法模型集 M。

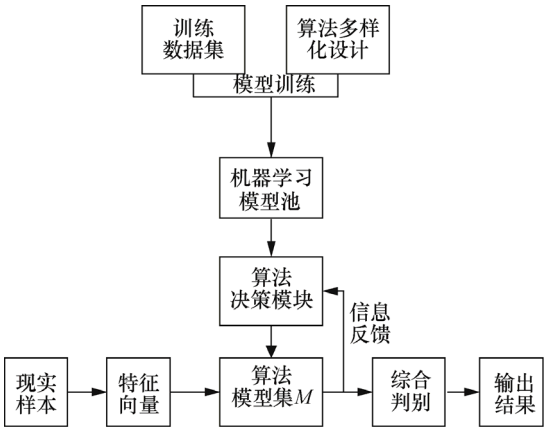


图 1 基于 MTD 技术的动态算法系统流程

Figure 1 MTD-based dynamic algorithm workflow

5 实验验证

本文使用对便携式文档格式（PDF，portable document format）的恶意性检测任务检验 MTD 技术对算法稳健性的提升效果。由于 PDF 文件的广泛应用和其存在的大量漏洞[20]，该格式文件越来越多地被用于发起 APT 攻击。虽然许多包括机器学习在内的检测算法已经被提出，但面对攻击者对机器学习模型本身的攻击时，其分类的准确率迅速下降。本文从 MTD 的角度分别对算法选择、特征选择和算法输出 3 方面进行多样化设计，并测试多样化的算法模型对对抗样本的检测能力。用来训练和测试的恶意样本集合 Mal 来自 VirusTotal，其收集了大量已知的恶意 PDF 样本；良性的样本集 Ben 来自互联网的收集，其均已通过多种反病毒系统的检测。其中，Mal 包含 10 986 个恶意样本，Ben 包含 8 969 个良性样本。为了使用机器学习模型对 PDF 文档进行检测，从结构和内容两个方面对 PDF 文件进行特征提取，并生成 296 维列向量，作为样本的特征空间。

5.1 算法选择多样性稳健性评估

抽取 Mal 和 Ben 各 5 000 个样本构成的平衡数据集训练支持向量机、决策树、逻辑回归分类模型，分别为 SVM、DT 和 LR。此实验方案下，配置 d 中 name=model，value 的取值范围为 {SVM,DT,LR}，配置集 D={d}，算法模型 m=((model,value))，模型集 M={(model, SVM), (model, DT),(model, LR)}，约束集 C 设定为对各模型准确率的要求不低于 90%。鉴于实验仅对算

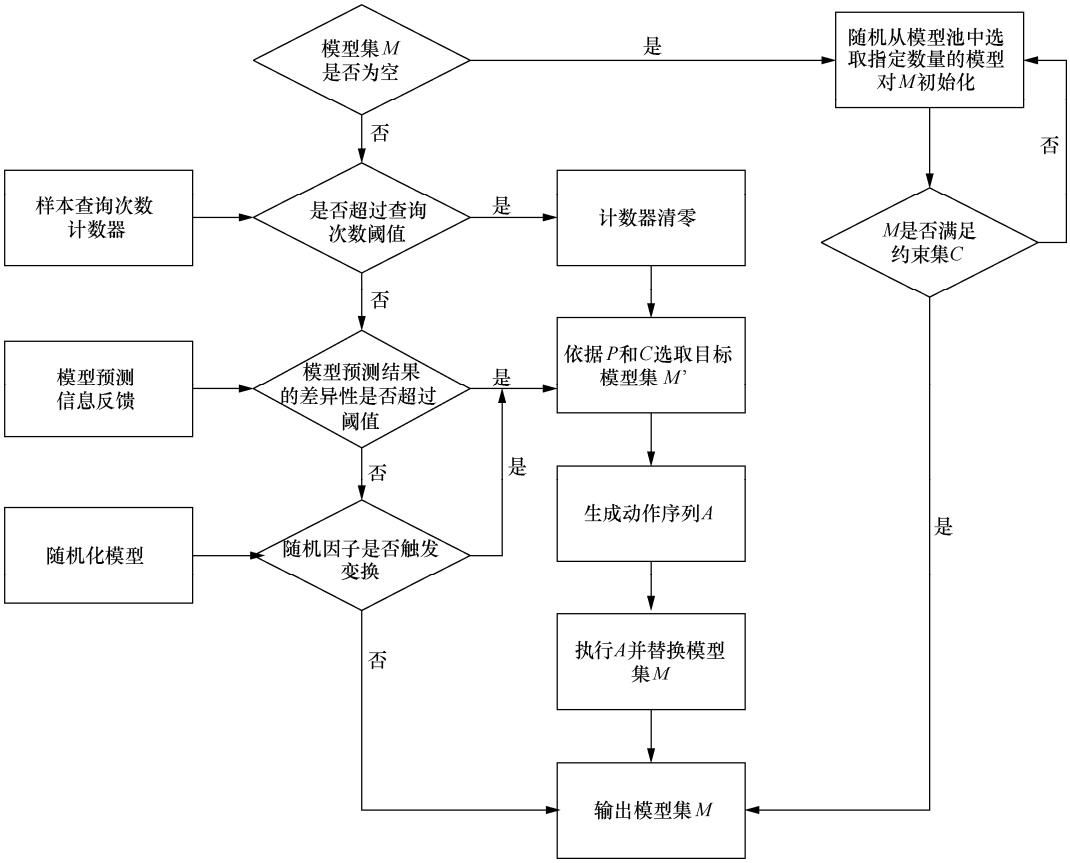


图 2　算法决策模块的工作流程
Figure 2　Algorithm decision module workflow

法的稳健性进行评估，策略集、动作集以及随机化方法集均可在构建完整的算法系统后再行决定。使用零阶优化算法[21]和梯度下降算法[22]针对上述模型展开攻击并生成对抗样本集 S_{adv}、D_{adv} 和 L_{adv}，这 3 个样本集均能欺骗各自的分类模型使其分类结果为良性。3 个机器学习模型对各数据集的分类结果如表 1 所示。

实验使用 Mal 和 Ben 中 $\frac{1}{2}$ 的样本训练 3 个机器学习模型，余下的 $\frac{1}{2}$ 为测试集。实验表明 3 个模型表现出良好的泛化能力，准确率高于 97%。各模型面对自己为目标的对抗样本集时毫无防御能力，但可以部分检测出其他模型的对抗样本。支持向量机模型防御对抗样本的能力较弱，而决策树算法能够防御绝大部分对抗样本，逻辑回归算法的防御能力介于二者之间。当攻击者无法察觉目标分类算法的切换时，其生成的对抗样本有较高概率被动态变换后的算法拒绝，说明算法选择的多样性能够提升算法系统的稳健性。

表 1　3 个机器学习模型对各数据集的分类准确率
Table 1　The classification accuracy rates of three machine learning models on each dataset

样本集	支持向量机	决策树	逻辑回归
Mal	97.98%	99.97%	98.54%
Ben	99.94%	99.87%	99.86%
S_{adv}	0.00%	96.98%	56.00%
D_{adv}	6.57%	0.00%	93.04%
L_{adv}	16.63%	97.44%	0.00%

5.2　特征选择多样性的稳健性评估

在黑盒攻击中，一方面，攻击者只能通过把数据输入算法模型中查看输出获取信息，很难了解模型使用的特征；另一方面，商用的检测算法模型使用的特征信息是保密的。因而可以假设攻击者只能通过经验进行特征提取和特征选择。为了测试特征选择多样性对算法稳健性的影响，在

上述机器学习算法所使用的 296 个特征中随机选取 30 个作为支持向量机模型所使用的特征进行分类，重复 20 次，生成使用不同特征的多样化算法集 $M=\{M_1,M_2,\cdots,M_{20}\}$。此时配置 d 中 name=features， value 的取值范围为 $F=\{f_1,f_2,\cdots,f_n\}$，$n=C_{296}^{30}$， f_i 为对特征集的具体抽样。配置集 $D=\{d\}$， 算法模型 m=(features, value)， 模型集 $M=\{(\text{features},f_1),(\text{features},f_2),\cdots,(\text{features},f_{20})\}$， 约束集 C 设定为对各模型准确率的要求不低于 80%。在测试集上恶意样本的检测率如表 2 所示，各模型在测试集上均有不错的检测性能。

为了检验特征多样性对算法系统稳健性的影响，实验分别对每个模型 M_i 生成对抗样本集 $S_{\text{adv}-i}$，然后使用模型 M_j ($j\neq i$)对 $S_{\text{adv}-j}$ 检测，检测率如表 3 所示，其中，行序号为模型序号，列序号为相应的对抗样本集序号，交叉处为模型 M_i 对 $S_{\text{adv}-j}$ 的检测率。结果显示在每一行中，当 $i=j$ 时检出率最低，说明算法模型检测自身对抗样本集的能力最差，而使用其他特征训练的机器学习模型对该对抗样本集的检测率明显高。在攻击者欠缺算法所使用的特征知识时，多样化的样本特征选择能够有效提升针对特定模型的对抗样本的检测率，提升了算法系统的稳健性。

表 2 使用不同特征的 SVM 模型的恶意样本检测率

Table 2 Detection rates of malicious samples by SVM models with different features

模型	准确率	模型	准确率
1	99.36%	11	99.19%
2	98.81%	12	89.12%
3	84.30%	13	88.12%
4	98.75%	14	86.15%
5	86.81%	15	87.72%
6	98.30%	16	82.45%
7	99.26%	17	90.40%
8	97.10%	18	90.58%
9	82.80%	19	83.68%
10	91.95%	20	84.65%

5.3 算法输出多样性的稳健性评估

一些机器学习模型除给出检测结果外，还会给出样本被判定为良性和恶意的概率，这样给攻击者更多的信息以训练替代模型。算法输出的多样性通过对输出添加随机扰动增加攻击者基于梯度下降的方法生成对抗样本的开销。扰动应控制在合理的范围内以保证不影响高概率良性和恶意样本的判别。实验首先训练攻击者的目标模型，训练集通过该算法模型的判别得到属于良性或恶意的隶属度概率分类结果；然后对该概率进行一定范围内的扰动，攻击者使用扰动后的输出结果训练

表 3 使用不同特征的 SVM 模型对敌对样本的检测率

Table 3 Detection rates of adversarial samples by SVM models with different features

对抗样本集序号	模型序号																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.63%	0.92%	64.25%	98.71%	86.20%	96.94%	99.26%	16.80%	81.92%	82.97%	98.58%	42.92%	88.12%	66.15%	87.73%	82.44%	27.42%	90.58%	4.82%	76.49%
2	1.20%	1.19%	55.81%	98.72%	86.08%	98.30%	99.25%	15.59%	72.96%	92.00%	95.59%	89.12%	15.16%	66.11%	27.89%	82.45%	90.40%	90.58%	71.14%	76.86%
3	99.36%	80.36%	22.58%	98.72%	85.23%	85.72%	37.44%	41.57%	82.80%	87.12%	35.45%	87.06%	33.73%	65.93%	54.76%	82.45%	89.79%	90.58%	83.68%	77.92%
4	99.28%	91.3%	55.32%	1.20%	86.81%	98.36%	1.23%	18.07%	82.80%	86.82%	97.62%	89.12%	81.85%	66.15%	87.20%	82.37%	1.80%	90.58%	75.90%	76.53%
5	20.88%	22.07%	56.72%	98.75%	18.31%	98.54%	99.26%	26.24%	82.11%	96.67%	99.19%	89.12%	80.40%	65.86%	87.66%	52.94%	97.92%	90.58%	73.63%	78.96%
6	99.36%	91.36%	50.28%	98.75%	86.75%	2.29%	2.88%	30.69%	88.04%	85.54%	98.59%	89.12%	88.12%	66.15%	87.73%	82.58%	89.82%	84.39%	83.68%	87.02%
7	99.36%	91.28%	18.63%	53.83%	86.81%	1.77%	0.73%	89.76%	82.74%	88.66%	89.90%	89.07%	89.43%	62.30%	87.43%	82.37%	89.81%	90.58%	75.84%	84.65%
8	2.91%	5.43%	6.48%	13.98%	84.57%	84.96%	99.26%	2.83%	82.37%	86.87%	99.19%	7.50%	23.09%	66.15%	87.21%	82.45%	5.50%	90.58%	71.34%	78.91%
9	22.09%	22.16%	63.89%	98.75%	85.46%	99.30%	98.81%	28.62%	18.44%	89.44%	22.22%	85.98%	76.36%	65.89%	86.28%	82.16%	21.88%	30.14%	77.60%	77.78%
10	8.04%	98.81%	64.16%	98.72%	88.16%	99.50%	99.26%	58.65%	82.80%	7.50%	15.48%	87.34%	23.47%	66.15%	87.8%	82.45%	10.25%	84.79%	71.94%	76.8%
11	99.36%	98.06%	3.59%	98.72%	86.81%	97.55%	84.87%	97.10%	81.92%	8.12%	0.79%	89.12%	8.34%	66.15%	87.73%	82.45%	90.35%	90.58%	75.92%	75.79%
12	10.83%	92.09%	17.79%	98.75%	86.81%	98.42%	99.26%	17.24%	72.42%	18.70%	99.19%	9.66%	88.12%	65.86%	86.58%	82.45%	90.00%	90.58%	83.68%	76.19%
13	99.36%	13.28%	13.45%	98.81%	86.17%	98.30%	99.82%	62.91%	82.04%	78.19%	12.00%	89.12%	10.60%	66.15%	25.4%	82.45%	89.8%	90.58%	72.45%	84.63%
14	99.36%	93.83%	57.98%	98.75%	89.49%	98.36%	33.55%	45.58%	82.75%	92.03%	99.19%	88.77%	88.12%	22.23%	87.52%	82.41%	90.4%	90.58%	75.18%	84.65%
15	99.36%	85.32%	58.91%	98.72%	86.77%	98.30%	98.79%	60.57%	52.80%	78.32%	99.19%	86.25%	19.99%	66.12%	10.96%	82.37%	89.97%	90.58%	64.67%	24.30%
16	99.36%	98.81%	63.94%	98.72%	88.42%	99.12%	98.87%	92.33%	82.67%	91.16%	99.19%	89.12%	88.12%	66.12%	87.38%	30.14%	89.90%	90.58%	83.68%	79.58%
17	99.28%	98.81%	32.43%	16.31%	87.28%	93.77%	98.79%	24.54%	82.68%	86.01%	98.68%	84.54%	79.17%	66.15%	87.62%	82.45%	8.65%	90.58%	76.62%	77.24%
18	99.36%	92.8%	57.21%	98.75%	86.80%	89.80%	99.26%	91.24%	22.64%	82.91%	98.75%	89.12%	88.12%	65.89%	87.72%	82.55%	89.85%	20.00%	77.65%	84.61%
19	16.61%	90.47%	63.84%	98.71%	86.19%	97.70%	17.03%	62.08%	82.78%	23.27%	90.78%	89.12%	74.98%	64.20%	80.34%	82.45%	89.81%	90.58%	13.83%	77.17%
20	99.28%	92.35%	15.26%	98.71%	86.80%	99.47%	99.26%	24.51%	82.53%	26.54%	90.51%	84.28%	88.12%	66.15%	30.19%	82.25%	89.51%	90.58%	89.60%	13.03%

替代模型并生成对抗样本。实验抽取良性样本和恶意样本集中各 5 000 个训练目标模型，并用整个数据集（10 986 个良性样本和 8 969 个恶意样本）做性能测试，其分类性能的混淆矩阵如表 4 所示，统计分类模型把良性样本正确分类为良性、错误分类为恶意和把恶意样本错误分类为良性和错误分类为恶意的样本数目并展示在一个表格中成为混淆矩阵。实验对样本施加扰动时采取的策略为对分类概率添加扰动最大界限范围内产生均匀分布的扰动。此时配置 d 中 name=noise，value 的取值范围为 $A=\{0.05, 0.10,\cdots,0.45\}$，取值为随机扰动的上界。配置集 $D=\{d\}$，算法模型 m=(noise,value)，模型集 M={(noise,0.05), (noise,0.10),⋯,(noise,0.45)}，约束集 C 设定为对训练集分类结果的干扰不超过 10%，如图 3 所示，随着扰动最大范围的增加，样本分类结果扰动后发生变化的数目逐步增加。使用扰动后的实验数据训练攻击者的替代模型 M_s，然后针对 M_s 生成对抗样本集 $S_{\text{adv}-s}$。目标模型对 $S_{\text{adv}-s}$ 的检测结果如图 4 所示，随着概率最大扰动范围的增大，针对替代模型生成的敌对样本被原始的目标模型检测出的比例逐步增大，当扰动概率界限高达 0.45 时，仅对 528 个非对抗样本的输出结果造成干扰，约占测试样本数的 6%，但对约 13%的对抗样本产生干扰，具有显著差异性。

表 4　目标 SVM 模型性能的混淆矩阵
Table 4　Confusion matrix of the target SVM model

分类	标签为恶意	标签为良性
恶意	10 961	0
良性	25	8 969

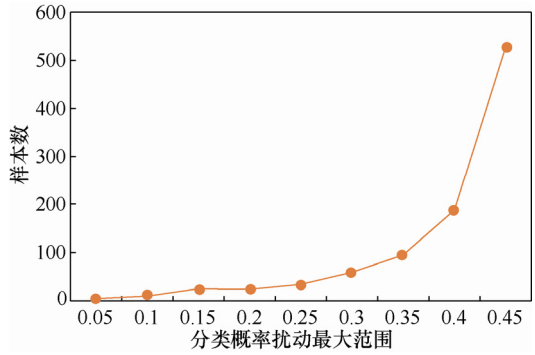


图 3　样本数随分类概率扰动最大范围的变化
Figure 3　The change of sample number with the maximum perturbation range of classification probability

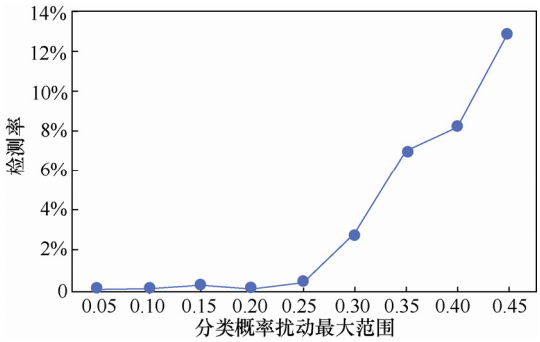


图 4　检测率随分类概率扰动最大范围的变化
Figure 4　The change of detection rate with the maximum perturbation range of classification probability

实验表明，算法输出的多样性能小幅度提高算法系统检测对抗样本的能力。

6　结束语

随着人工智能技术在各领域的广泛应用和蓬勃发展，机器学习起到越来越重要的作用。在提高算法分类的准确率、召回率的同时，提高机器学习算法面对敌对攻击时的稳健性也越来越重要，特别是在攻防对抗十分激烈的网络安全领域。为了解决上述问题，本文从 MTD 思想的角度对传统的算法设计流程进行改造。首先对该思想的作用进行形式化的定性分析；其次根据算法实际设计经验提出了 3 种算法的动态化设计；然后结合 MTD 技术原理提出了动态算法的工作流程；最后以恶意 PDF 文档检测任务为例检验了 MTD 技术对算法稳健性的增益。实验证明，3 种方法均能不同程度地提高算法的稳健性。本文设计的基于 MTD 防御技术的动态算法系统不是替换现有的算法稳健性增强算法，而是在既有的重训练、蒸馏等方法的基础上做进一步的改造，是从新的角度对现有算法稳健性的增强。

参考文献：

[1] JIANG H, NAGRA J, AHAMMAD P. Sok: applying machine learning in security-a survey[J]. arXiv preprint arXiv:1611.03186, 2016.

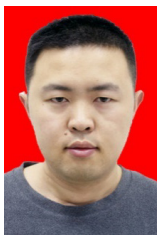
[2] PITROPAKIS N, PANAOUSIS E, GIANNETSOS T, et al. A taxonomy and survey of attacks against machine learning[J]. Computer Science Review, 2019, 34: 100199.

[3] 张东, 张尧, 刘刚, 等. 基于机器学习算法的主机恶意代码检测技术研究[J]. 网络与信息安全学报, 2017, 3(7): 25-32.
ZHANG D, ZHANG Y, LIU G, et al. Research on host malcode detection using machine learning[J]. Chinese Journal of Network and Information Security, 2017, 3(7): 25-32.

[4] 张骁敏, 刘静, 庄俊玺, 等. 基于权限与行为的 Android 恶意软件检测研究[J]. 网络与信息安全学报, 2017, 3(3): 51-57.

ZHANG X M, LIU J, ZHUANG J X, et al. Research on Android malware detection based on permission and behavior[J]. Chinese Journal of Network and Information Security, 2017, 3(3): 51-57.
[5] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013
[6] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv:1412.6572, 2014.
[7] ZHANG G, YAN C, JI X, et al. Dolphinattack: inaudible voice commands[C]//The 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 103-117.
[8] GROSSE K, PAPERNOT N, MANOHARAN P, et al. Adversarial perturbations against deep neural networks for malware classification[J]. arXiv preprint arXiv:1606.04435, 2016.
[9] CHEN S, XUE M, FAN L, et al. Automated poisoning attacks and defenses in malware detection systems: an adversarial machine learning approach[J]. Computers & Security, 2018, 73: 326-344.
[10] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//2016 IEEE Symposium on Security and Privacy (SP). 2016: 582-597.
[11] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (SP). 2017: 39-57.
[12] 蔡桂林, 王宝生, 王天佐, 等. 移动目标防御技术研究进展[J]. 计算机研究与发展, 2016, 53(5): 968-987.
CAI G L, WANG B S, WANG T Z, et al. Research and development of moving target defense technology[J]. Journal of Computer Research and Development, 2016, 53(5): 968-987.
[13] EVANS D, NGUYEN-TUONG A, KNIGHT J. Effectiveness of moving target defenses[M]//Moving Target Defense. 2011: 29-48.
[14] JAFARIAN J H, AL-SHAER E, DUAN Q. Openflow random host mutation: transparent moving target defense using software defined networking[C]//The First Workshop on Hot Topics in Software Defined Networks. 2012: 127-132.
[15] SENGUPTA S, CHAKRABORTI T, KAMBHAMPATI S. MTDeep: boosting the security of deep neural nets against adversarial attacks with moving target defense[C]//Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[16] LEI C, MA D H, ZHANG H Q. Optimal strategy selection for moving target defense based on Markov game[J]. IEEE Access, 2017, 5: 156-169.
[17] ROY A, CHHABRA A, KAMHOUA C A, et al. A moving target defense against adversarial machine learning[C]//The 4th ACM/IEEE Symposium on Edge Computing. 2019: 383-388.
[18] 李亚龙, 陈勤, 张旻. 基于博弈论的移动目标最优防御策略研究[J]. 计算机工程与应用, 2019, 55(19): 141-146.
LI Y L, CHEN Q, ZHANG M. Research on optimal defense strategy of moving targets based on game theory[J]. Computer Engineering and Applications, 2019, 55(19): 141-146.
[19] KANTCHELIAN A, TYGAR J D, JOSEPH A. Evasion and hardening of tree ensemble classifiers[C]//International Conference on Machine Learning. 2016: 2387-2396.
[20] NISSIM N, COHEN A, GLEZER C, et al. Detection of malicious PDF files and directions for enhancements: a state-of-the art survey[J]. Computers & Security, 2015, 48: 246-266.
[21] CHEN P Y, ZHANG H, SHARMA Y, et al. Zoo: zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]//The 10th ACM Workshop on Artificial Intelligence and Security. 2017: 15-26.
[22] MUÑOZ-GONZÁLEZ L, BIGGIO B, DEMONTIS A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization[C]//The 10th ACM Workshop on Artificial Intelligence and Security. 2017: 27-38.

[作者简介]



何康（1992− ），男，山东济宁人，信息工程大学博士生，主要研究方向为网络空间安全。

祝跃飞（1962− ），男，河南郑州人，信息工程大学教授、博士生导师，主要研究方向为入侵检测、密码学、信息安全。

刘龙（1983− ），男，河南郑州人，信息工程大学讲师，主要研究方向为入侵检测和信息安全。

芦斌（1983− ），男，河南郑州人，信息工程大学副教授，主要研究方向为信息安全、机器学习和网络分析。

刘彬（1981− ），女，河南郑州人，信息工程大学副教授，主要研究方向为网络安全。