对抗环境下基于集成决策树的恶意 PDF 文件检测

李坤明[1] 顾益军[1] 张培晶[2]

[1](中国人民公安大学信息技术与网络安全学院 北京 102600)

[2](中国人民公安大学网络信息中心 北京 102600)

摘 要 决策树算法在恶意 PDF 文件检测上具有较高的精确度。当存在攻击者通过探索模型的漏洞时,会使恶意 PDF 样本轻易逃避检测,导致模型的鲁棒性较差。针对该问题,提出一种集成决策树的方法以提升模型的鲁棒性。将攻击产生的对抗样本添加到训练集中;使用 Adaboost 方法集成决策树构建分类模型;将训练后的模型与现有的集成方法进行比较。实验结果表明:该方法在检测恶意 PDF 文件分类中在保持较高分类精度的同时也具有更强的鲁棒性。

关键词 决策树算法 逃避攻击 集成学习 鲁棒性 Adaboost

中图分类号 TP393 **文献标志码** A DOI: 10.3969/j.issn.1000-386x.2020.10.051

DETECTION OF MALICIOUS PDF FILES BASED ON INTEGRATED DECISION TREE IN ADVERSARIAL ENVIRONMENT

Li Kunming[1] Gu Yijun[1] Zhang Peijing[2]

[1](*College of Information Technology and Network Security, People's Public Security University of China, Beijing 102600, China*)

[2](*Network Information Center, People's Public Security University of China, Beijing 102600, China*)

Abstract Decision tree algorithm has a high accuracy in malicious PDF file detection. When an attacker explores the vulnerabilities of the model, the malicious PDF samples can easily escape detection, resulting in poor robustness of the model. To solve this problem, a method of integrated decision trees is proposed to improve the robustness of the classification model. The adversarial samples generated by the attack were added to the training set, and the Adaboost method was used to build the decision tree to construct the classification model. We compared the trained model with the existing ensemble method. The experimental results show that the proposed method has higher classification accuracy and stronger robustness in detecting malicious PDF file classification.

Keywords Decision tree algorithm Attack evasion Ensemble learning Robustness Adaboost

0 引 言

由于 PDF 文件格式的稳定性和跨平台的交互性,其在政府、企业以及社会等组织的日常办公中被广泛使用。针对 PDF 文件漏洞发起的攻击迅速增多,对恶意 PDF 文件的检测非常有必要[1]。随着人工智能的迅速发展,机器学习方法在恶意 PDF 文件的检测中有着广泛的应用。初始对于恶意 PDF 文件检测的研究方法主要是通过定位和抽取恶意 PDF 文件中的 JavaScript 代码进行检测[2-3],但是由于大量的 JavaScript 代码通过加密、混淆等手段使得在对其进行定位和解析中存在着巨大困难,此外还存在一些恶意 PDF 文件不含有 JavaScript 代码。这些因素都使得使用该方法进行恶意 PDF 文件检测时因为重要特征的缺失导致检测的效率较低。于是出现了另一种通过将恶意 PDF 文件中元数据作为特征进行检测的方法,其中较为常用的是将恶意 PDF 文件中的关键词[4]、结构路径[5-6]

收稿日期:2019-07-05。国家重点研发计划项目(2017YFC0820100)。李坤明,硕士生,主研领域:机器学习,网络安全。顾益军,教授。张培晶,讲师。

作为特征进行检测,结果表明该方法具有较高的检测率,由原来的 80% 提升到 90% 以上。

研究发现通过使用元数据作为特征进行恶意 PDF 文件检测时,存在着一些非法攻击者通过精心构造恶意 PDF 样本(对抗样本)使其逃避分类模型的检测,降低了分类模型的可用性,导致模型的鲁棒性较差。因此,研究对抗环境下在保持分类模型对恶意 PDF 文件检测具有较高检测率的同时也具有较强的鲁棒性具有重要意义。

1　相关工作

1.1　逃避攻击

目前机器学习方法在对文本进行分类时面临的攻击问题根据发生的时间可以分为两类,分别是针对训练阶段的攻击和针对测试阶段的攻击。本文只考虑在测试阶段发生的攻击。

逃避攻击是发生在机器学习模型测试阶段的一种常用的攻击方式[7]。其原理是在基于特征提取的分类和恶意软件检测中,攻击者能够通过一定的方式修改恶意样本中的部分特征进而生成对抗样本,对抗样本本身仍保留恶意内容,但使得其被模型误分类为正常样本,逃避了模型的检测进而降低了模型的可用性。对于一个恶意 PDF 样本,在无攻击时其输出为(x,y),在对抗环境下,攻击者通过一定的方式修改其特征使得输出改变为(x^*,z),表示如下:

$$\Delta(x,z)=\inf_{x^*}\{\|x^*-x\|:f(x^*)=z\} \tag{1}$$

其中最优的攻击策略是在使得分类模型误分类的同时,所需要修改的特征值最小,C 为标签集合,即:

$$\Delta(x)=\min_{z\subset C_n\{y\}}\Delta(x,z) \tag{2}$$

针对恶意 PDF 文件检测的分类模型中,逃避攻击方法产生对抗样本的方式通常有两种,一种是基于梯度的方法产生对抗样本[8],另一种是基于特征加法的方式产生对抗样本[9]。这两种方法在垃圾邮件检测系统与恶意 PDF 文件检测上都实现了攻击。由于攻击者实施攻击不需要了解训练数据集的分布,并且能够在已知更少的信息下实现对分类模型的攻击,因此逃避攻击是威胁机器学习安全的一种重要攻击方式。

1.2　对抗环境下机器学习防御技术

机器学习方法被广泛应用到网络入侵检测、人脸识别、文本分类等安全领域。机器学习模型经常受到对抗样本的干扰,在利益的驱使下,攻击者通过构造对抗样本恶意地干扰输入,以便在测试时使得分类模型误分类[10]。

在对抗环境下对机器学习鲁棒性的防御技术主要有蒸馏法、正则化方法、对抗性训练方法和重新构造分类模型等。Papernot 等[11]提出了蒸馏防御机制,该方法通过提升攻击者产生对抗样本时需要修改特征最小平均值的方式提高模型的鲁棒性。但随后 Carlini 等[12]指出该方法在不知道分类模型时的效果不明显。文献[13-14]使用正则化和对抗性训练的方式,通过专门的训练增强模型的鲁棒性,但仍然存在盲区。针对如何在测试阶段构建一个鲁棒性能好的分类模型主要有以下研究: Zhang 等[15]通过使用封装式(Wrapper)的特征选择方法挑选部分特征子集的方式构建出鲁棒的分类模型; Bhagoji 等[16]通过使用主成分分析方式对特征进行降维,构建了一个鲁棒性较好的分类模型; Biggio 等[17]提出使用 Bagging 集成的方法构建出鲁棒的分类模型。

目前使用集成学习的方式对抗逃避攻击通常是将若干个单分类器并行叠加在一起,然后将多个单分类器的结果以投票法、均值法等结合策略输出,这些提高模型鲁棒性的方法没有考虑到攻击者产生的对抗样本信息。本文提出一种新的集成方法,首先通过模拟攻击者的攻击将攻击过程中产生的对抗样本添加到训练集中; 然后使用 Adaboost 方法集成决策树在每次迭代过程中增加错分样本的权重,构建出新的恶意 PDF 文件检测模型; 最后通过模拟不同强度的攻击测试所提方法在恶意 PDF 文件检测上的有效性,并与单分类器决策树方法、Bagging 集成方法相比较。实验结果表明,本文方法在无攻击情况下具有较高的分类精度,有攻击时的鲁棒性优于其他两种方法。

2　存在的问题

2.1　基于决策树的恶意 PDF 文件检测的脆弱性

PDF 格式的灵活性使得攻击者有足够的机会改变其内容与结构。在逃避攻击中攻击者的目的是通过修改恶意 PDF 文件使其被分类模型误分为正常样本。由于 PDF 文件格式的特性使得攻击者很难删除恶意 PDF 文件中的部分恶意内容,但是可以轻松地向恶意 PDF 文件中添加正常内容,因此在恶意 PDF 文件检测中常用基于特征加法的攻击方式。

决策树算法应用于恶意 PDF 文件检测时,攻击者通过查找终节点被正确分为恶意的路径,然后沿着这条路径回溯并找到第一个非终节点,沿着这个方向的终节点被分为正常样本。在这种情况下,攻击者通过

向恶意样本中添加正常样本含有的特征生成对抗样本,改变决策树分类模型的决策路径进而导致恶意样本被误分类,使得模型的鲁棒性较差。

2.2 基于集成决策树的方法提升模型的鲁棒性

由于 JavaScript 代码具有隐蔽性,对于其在 PDF 文件中的定位比较困难,因此使用 JavaScript 代码的检测率较低。在随后提出的基于结构特征的恶意 PDF 文件检测方法中,文献[5]提出了基于结构特征的 PDF 文件特征抽取方法,并使用决策树算法对恶意 PDF 文件进行分类,实验结果表明该方法具有较高的正确率。但该方法存在两个问题:一是文档特征复杂,仅以特征在每个文档中出现的频数作为特征值不够客观;二是在对抗环境下,存在非法的攻击者通过修改测试集中恶意 PDF 文件中的特征来逃避分类器的检测,表明该模型的鲁棒性较差。

对此提出一种基于集成决策树的恶意 PDF 文件检测方法。在构建分类模型前就考虑到攻击者的攻击问题,通过模拟针对于恶意 PDF 文件的特征加法攻击,将攻击产生的对抗样本添加到训练集中。然后使用 Adaboost 集成决策树的方法构建分类模型,因为攻击者产生的对抗样本会被分类模型错分为正常样本,Adaboost 方法使用串行迭代的方式,在每一次训练中会增加错分样本的权重,通过模拟攻击产生的对抗样本的权重则会在下一次训练中增加。因此攻击者再对新的分类模型实施攻击时,达到相同的攻击效果需要修改恶意 PDF 文件的最小平均特征值会增加。最后通过实验将该方法与单分类模型和现有的 Bagging 集成方法进行比较,结果表明在分类精度与鲁棒性上均高于这两种方法。

2.3 对抗环境下分类器性能评价指标

传统的分类模型通常只考虑分类器的分类精确度,没有考虑到攻击者的攻击问题。存在非法攻击者通过修改恶意样本的特征,使其被分类器检测为正常样本。因此在评价分类器性能时不仅要考虑分类模型在无攻击时的分类精确度,还要考虑到其本身的鲁棒性,即对抗逃避攻击的能力。对于恶意 PDF 文件检测的二分类问题,其分类结果的混淆矩阵如表 1 所示。

表 1 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP(假反例)	FN(假反例)
反例	FP(假反例)	TN(真反例)

在分类模型的评价指标中,正确率表示模型检测结果是正确的 PDF 文件数占 PDF 文件总量的比例。逃避攻击中,攻击者的目的在于修改恶意样本使其被模型误分类,漏报率表示恶意 PDF 文件被模型检测为正常 PDF 文件的数量占恶意 PDF 文件总量的比例,是评价分类模型鲁棒性的重要指标。因此本文选择精确度(Acc)和漏报率(FNR)作为评价分类模型的指标。

$$Acc = \frac{TP + TN}{TP + TN + FN + FP} \tag{3}$$

$$FNR = \frac{FN}{FN + TP} \tag{4}$$

3 模型设计

在对抗环境下,为提高分类模型对恶意 PDF 文件的检测率以及模型的鲁棒性。本文提出一种集成决策树的方法检测恶意 PDF 文件,具体步骤如下:

1) 对于一个 PDF 文件,具有固定的格式,通过使用 PDF 文件解析器可以抽取出其结构特征,恶意 PDF 文件数据集可以表示为 $T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$,$x_i = (x_{i1}, x_{i2}, \cdots, x_{ij})$。其中:$x_i$ 表示一个空间维度为 j 的 PDF 文件;$y_n \in \{1, -1\}$ 为 PDF 文件的标签,1 表示恶意文件,-1 表示正常文件。

2) 使用 TF-IDF(Term Frequency-Inverse Document Frequency)算法对每个 PDF 文件所抽取的特征属性进行量化处理,计算每一个特征向量的权重。其公式如下:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{5}$$

$$idf_i = \log \frac{|D|}{1 + |\{j: t_i \in d_j\}|} \tag{6}$$

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{7}$$

式中:$n_{i,j}$是特征 t_i 在一个 $|\{j: t_i \in d_j\}|$ PDF 文件 d_j 中出现的次数;$\sum_k n_{k,j}$ 是所有特征出现的次数之和;$|D|$表示训练集中 PDF 文件的数量;$|\{j: t_i \in d_j\}|$表示包含特征 t_i 的 PDF 文件数量。

3) 使用基于特征加法攻击方法实现对决策树算法的攻击,并记录分类器的分类效果。

4) 使用 Adaboost 方法集成决策树算法,在每一次串行迭代训练中,增加攻击过程中错分样本的权重,经过 M 次迭代训练出新的分类模型。对于训练数据集 $D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$,$x_i \in X \subseteq \mathbf{R}$,$y_i \in$

$\{-1,1\}$。

(1) 初始化训练数据的权值分布:

$$D_1=(w_{11},w_{12},\cdots,w_{1N})\quad w_{1i}=\frac{1}{N},i=1,2,\cdots,N \tag{8}$$

(2) 对 $m=1,2,\cdots,M$(M 表示实验设置决策树的个数):

① 使用具有权值分布的 D_m 训练数据集学习,得到决策树模型 $G_m(x)$。

② 计算 $G_m(x)$ 在训练数据集上的分类误差率:

$$e_m=\sum_{i=1}^{N}w_{mi}I(G_m(x_i)\neq y_i) \tag{9}$$

③ 计算得到决策树模型 $G_m(x)$ 的权重:

$$a_m=\ln(1-e_m)/e_m \tag{10}$$

④ 更新 PDF 训练数据集的权值分布:

$$w_{m+1,i}=\frac{w_{mi}}{Z_m}\exp(-a_m y_i G_m(x_i)) \tag{11}$$

式中: Z_m 为规范化因子。

(3) 得到由决策树集成的分类器:

$$F(x)=\operatorname{sign}\left(\sum_{i=1}^{n}\alpha_m G_m(x)\right) \tag{12}$$

5) 通过模拟不同强度的攻击,对集成决策树模型 $F(x)$ 进行攻击,验证本文方法构建的恶意 PDF 文件检测模型在逃避攻击情况下的分类效果。具体检测流程如图 1 所示。

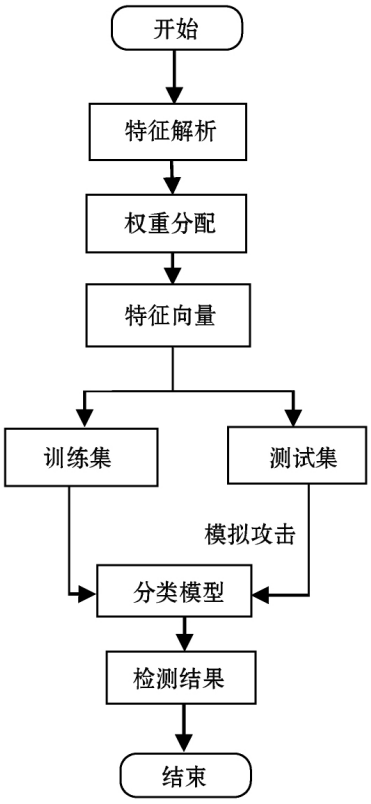

图 1　集成决策树算法的恶意 PDF 文件检测流程

4　实验分析

4.1　实验数据

本文使用的数据集来源于 CONTAGIO 数据集[18],选取了其中的 4 786 个恶意 PDF 文件和 4 904 个正常 PDF 文件,使用基于关键词的方法提取 PDF 文件特征。采取 5 折交叉验证的方式进行实验,即将数据集随机分成 5 份,其中 4 份作为训练集,剩余的 1 份作为测试集。

4.2　攻击强度

本文方法旨在提高分类模型在测试过程中分类模型对抗逃避攻击的能力,实验采用基于特征加法的攻击方式。基于特征加法的攻击方法是机器学习在文本分类中一种常见的攻击方式,同时也是检测分类模型鲁棒性的方式。本文将攻击强度记为 K,通过设定 K 为 2、4、6、8、10,即向恶意样本中添加不同的正常样本特征属性的数量,模拟不同强度的攻击。

4.3　实验结果

本文通过将集成决策树的个数 M 设置为 10 到 70,并测试集成方法在无攻击时的分类精确度。由表 2 可知,随着弱分类器个数的增加,分类效果并未提升,此外考虑到计算开销问题以及模型的泛化性能,本文选择集成决策树的个数为 10 进行实验。

表 2　无攻击时集成不同个数单分类器的精确度

方法	M						
	10	20	30	40	50	60	70
Adaboost-DT	0.981 6	0.981 1	0.981 6	0.981 1	0.981 1	0.981 2	0.981 1
Bagging-DT	0.981 5	0.981 1	0.981 2	0.981 4	0.981 4	0.981 0	0.981 2

由表 3 可知,在无攻击的情况下($K=0$),基于 Adaboost 方法集成决策树的方法在恶意 PDF 文件检测的精确度上相比 Bagging 方法增加了 0.56%,但两种集成的方法在检测精确度上均高于决策树算法构建的单分类器。随着攻击强度的增加,决策树分类器的检测精确度迅速下降,当攻击强度 $K=10$ 时,其恶意 PDF 文件检测精确度由初始的 97.01% 下降到 71.59%,下降了 25.42 个百分点。相同攻击强度下,基于 Bagging 的集成方法下降了 10.95 个百分点,本文的方法下降了 8.94 个百分点,说明本文方法在不同攻击强度下具有更高的检测精确度,对抗逃避攻击时模型的鲁棒性更好。

表 3 三种方法在不同攻击强度下的精确度

攻击强度 K	DT	Bagging-DT	Adaboost-DT
0	0.970 1	0.981 0	0.986 6
2	0.906 1	0.980 4	0.980 5
4	0.850 3	0.949 0	0.961 4
6	0.810 9	0.923 8	0.942 3
8	0.763 1	0.898 7	0.921 1
10	0.715 9	0.871 5	0.897 2

在漏报率方面，两种集成方法在无攻击时较为接近。随着攻击强度的增加，由表 4 可知，基于决策树算法的检测模型随着攻击强度的增加在漏报率上下降幅度最大。当攻击强度 $K = 10$ 时，基于决策树的恶意 PDF 文件检测模型的漏报率为 10%，基于 Bagging 集成的方法漏报率为 8.1%，使用本文方法的漏报率最低为 5.07%。

表 4 三种方法在不同攻击强度下的漏报率

攻击强度 K	DT	Bagging-DT	Adaboost-DT
0	0.009 0	0.000 8	0.000 4
2	0.031 0	0.009 8	0.001 7
4	0.049 9	0.038 8	0.030 0
6	0.068 5	0.059 9	0.032 1
8	0.090 5	0.076 7	0.049 5
10	0.100 0	0.081 0	0.050 7

向恶意 PDF 文件中添加正常样本的特征，会改变决策树分类模型划分属性选择，导致模型做出错误的判断，因此单分类器在面对攻击时模型的鲁棒性通常较差。Bagging 集成方法的原理是通过使用自助采样法并行训练出多个单决策树分类模型，使用投票的方法得到最终结果。本质上减小了每个单分类器输出结果的权重，虽然在一定程度上模型的分类鲁棒性优于单分类器，但是模型的构建过程中没有考虑到攻击者产生的逃避攻击样本问题。而本文采用的方法通过使用串行迭代的方式生成模型，并在训练集中加入对抗样本，在每一轮训练过程中会根据样本的分布为每个样本重新赋予一个权重，使得样本属性的权重分配更加均衡合理，即使存在攻击者通过修改恶意 PDF 文件特征逃避检测时，达到相同的攻击效果，其修改的特征值也就越多。因此本文方法构建的分类模型在恶意 PDF 文件检测上具有较高检测精确度的同时，模型的鲁棒性更好。

5 结 语

决策树算法在恶意 PDF 文件检测上通常具有较高的精确度，但模型的鲁棒性较差。为此，本文在训练分类模型的过程中将攻击者产生的对抗样本添加到训练集中，然后使用串行迭代的方式在每次迭代的过程中通过不断增大错分样本的权重，最终训练出分类模型。为验证本文构建模型的分类性能和鲁棒性，通过模拟不同强度的攻击与单分类器和 Bagging 集成的多分类器进行比较。实验结果表明，本文方法在无攻击情况下具有较高的精确度，同时在不同攻击强度下其精确度和漏报率均优于决策树算法和 Bagging 集成方法。

本文主要是对决策树算法的集成，构建了一个鲁棒的恶意 PDF 文件检测模型。下一步可以使用集成的方法将多种算法融合在恶意 PDF 文件的检测上，还可以考虑将集成的方法与其他提高模型鲁棒性的方法相结合进行研究。

参 考 文 献

[1] 杜学绘，林杨东，孙奕. 基于混合特征的恶意 PDF 文档检测[J]. 通信学报，2019，40(2)：118－128.

[2] 孙本阳，王轶骏，薛质. 一种改进的恶意 PDF 文档静态检测方案[J]. 计算机应用与软件，2016，33(3)：308－313.

[3] 冯迪. 恶意 PDF 文档检测技术研究与实现[D]. 哈尔滨：哈尔滨工程大学，2017.

[4] Maiorca D，Biggio B. Digital investigation of PDF files：unveiling traces of embedded malware[J]. IEEE Security & Privacy，2019，17(1)：63－71.

[5] Srndic N，Laskov P. Detection of malicious PDF files based on hierarchical document structure[C]//Proceedings of the 20th Annual Network & Distributed System Security Symposium，2013：1－16.

[6] Srndic N，Laskov P. Hidost：A static machine-learning-based detector of malicious files[J]. EURASIP Journal on Information Security，2016(1)：45.

[7] Tong L，Li B，Chen H，et al. How robust is robust ML? Evaluating models of classifier evasion in PDF malware detection[EB]. arXiv：1708.08327v4，2017.

[8] Biggio B，Corona I，Maiorca D，et al. Evasion attacks against machine learning at test time[C]//2013 Joint European Conference on Machine Learning and Knowledge Discovery in Databases(ECML PKDD). Springer，2013：387－402.

(下转第 333 页)

结果表明灰度直方图的随机森林、灰度共生矩阵的随机森林、N-Gram 的随机森林，以及融合特征的随机森林均可以有效地进行恶意代码的分类，其中 3 种特征融合后与随机森林算法相结合其分类效果显著提升。目前更多的是静态特征融合，从代码生成的灰度图纹理和恶意代码文本两方面为切入点，取得不错的分类效果。下一步将融合一些动态特征，比如恶意软件的行为特征，观察其是否可以进一步提高恶意代码分类的准确率，从而进一步优化分类器。

参 考 文 献

[1] 国家互联网应急中心(CNCERT). 2018 年我国互联网网络安全态势报告[OL]. [2019-04-18]. https://www.freebuf.com/articles/network/201280.html.

[2] 奇虎 360. 中国互联网安全报告[EB/OL]. 2018-08-02. https://www.freebuf.com/articles/paper/179295.html.

[3] 崔鸿雁，徐帅，张利锋，等. 机器学习中的特征选择方法研究及展望[J]. 北京邮电大学学报，2018，41(1)：1-9.

[4] 高程程，惠晓威. 基于灰度共生矩阵的纹理特征提取[J]. 计算机系统应用，2010，19(6)：195-198.

[5] 周绮凤，洪文财，杨帆，等. 基于随机森林相似度矩阵差异性的特征选择[J]. 华中科技大学学报(自然科学版)，2010，38(4)：58-61.

[6] 王卫红，朱雨辰. 基于 N-Gram 与加权分类器集成的恶意代码检测[J]. 浙江工业大学学报，2017，45(6)：604-632.

[7] Breiman L. Random forest[J]. Machine Learning，2001，45：5-32.

[8] Nataraj L，Karthikeyan S，Jacob G，et al. Malware images：visualization and automatic classification[C]//8th International Symposium on Visualization for Cyber Security. ACM，2011.

[9] 戴逸辉，殷旭东. 基于随机森林的恶意代码检测[J]. 网络空间安全，2018，9(2)：70-75.

[10] Kaggle[OL]. https://kaggle.com/c/maleware classification/.

(上接第 322 页)

[9] Rndic N，Laskov P. Practical evasion of a learning-based classifier：A case study[C]//2014 IEEE Symposium on Security and Privacy. IEEE，2014：197-211.

[10] 张思思，左信，刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报，2019，42(8)：1886-1904.

[11] 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). IEEE，2016：582-597.

[12] Carlini N，Wagner D. Defensive distillation is not robust to adversarial examples[EB]. arXiv：1607.04311，2016.

[13] Graese A，Rozsa A，Boult T E. Assessing threat of adversarial examples on deep neural networks[C]//2016 15th IEEE International Conference on Machine Learning and Applications(ICMLA). IEEE，2016：69-74.

[14] Shaham U，Yamada Y，Negahban S. Understanding adversarial training：Increasing local stability of supervised models through robust optimization[J]. Neurocomputing，2018，307：195-204.

[15] Zhang F，Chan P P，Biggio B，et al. Adversarial feature selection against evasion attacks[J]. IEEE Transactions on Cybernetics，2016，46(3)：766-777.

[16] Bhagoji A N，Cullina D，Mittal P. Dimensionality reduction as a defense against evasion attacks on machine learning classifiers[EB]. arXiv：1704.02654，2017.

[17] Biggio B，Corona I，Fumera G，et al. Bagging classifiers for fighting poisoning attacks in adversarial classification tasks[C]//International workshop on multiple classifier systems. Springer，2015：350-359.

[18] Xu W L，Qi Y J，Evans D. Automatically evading classifiers[C]//Proceedings of the 2016 Network and Distributed Systems Symposium，2016：21-24.

(上接第 327 页)

[12] Dauphin Y N，Fan A，Auli M，et al. Language modeling with gated convolutional networks[EB]. arXiv：1612.08083，2016.

[13] 冉鹏，王灵，李昕，等. 改进 Softmax 分类器的深度卷积神经网络及其在人脸识别中的应用[J]. 上海大学学报(自然科学版)，2018，24(3)：352-366.

[14] Krawczyk B. Learning from imbalanced data：open challenges and future directions[J]. Progress in Artificial Intelligence，2016，5(4)：221-232.

[15] Lin T Y，Goyal P，Girshick R，et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2018，42(2)：318-327.

[16] Simonyan K，Zisserman A. Very deep convolutional networks for large-scale image recognition[EB]. arXiv：1409.1556，2014.

[17] Kinga D P，Adam J B. Adam：A method for stochastic optimization[C]//International Conference on Learning Representations(ICLR)，2015.

[18] Tavallaee M，Bagheri E，Lu W，et al. A detailed analysis of the KDD CUP 99 data set[C]//2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE，2009：1-6.