基于粒子群优化的对抗样本生成算法

钱亚冠*[①] 卢红波[①] 纪守领[②] 周武杰[③] 吴淑慧[①]
云本胜[①] 陶祥兴[①] 雷景生[③]

[①](浙江科技学院理学院/大数据学院 杭州 310023)
[②](浙江大学计算机学院 杭州 310027)
[③](浙江科技学院电子与信息工程学院 杭州 310023)

摘 要：随着机器学习被广泛的应用，其安全脆弱性问题也突显出来。该文提出一种基于粒子群优化(PSO)的对抗样本生成算法，揭示支持向量机(SVM)可能存在的安全隐患。主要采用的攻击策略是篡改测试样本，生成对抗样本，达到欺骗SVM分类器，使其性能失效的目的。为此，结合SVM在高维特征空间的线性可分的特点，采用PSO方法寻找攻击显著性特征，再利用均分方法逆映射回原始输入空间，构建对抗样本。该方法充分利用了特征空间上线性模型上易寻优的特点，同时又利用了原始输入空间篡改数据的可解释性优点，使原本难解的优化问题得到实现。该文对2个公开数据集进行实验，实验结果表明，该方法通过不超过7%的小扰动量生成的对抗样本均能使SVM分类器失效，由此证明了SVM存在明显的安全脆弱性。

关键词：机器学习；支持向量机；探测攻击；显著性扰动；对抗样本

中图分类号：TP309.2 **文献标识码**：A **文章编号**：1009-5896(2019)07-1658-08

DOI: 10.11999/JEIT180777

Adversarial Example Generation Based on Particle Swarm Optimization

QIAN Yaguan[①] LU Hongbo[①] JI Shouling[②] ZHOU Wujie[③] WU Shuhui[①]
YUN Bensheng[①] TAO Xiangxing[①] LEI Jingsheng[③]

[①](*School of Science/School of Big-data Science, Zhejiang University of Science and Technology, Hangzhou* 310023, *China*)
[②](*School of Computer Science, Zhejiang University, Hangzhou* 310027, *China*)
[③](*School of Electronic and Information Engineering, Zhejiang University of Science and Technology, Hangzhou* 310023, *China*)

Abstract: As machine learning is widely applied to various domains, its security vulnerability is also highlighted. A PSO (Particle Swarm Optimization) based adversarial example generation algorithm is proposed to reveal the potential security risks of Support Vector Machine (SVM). The adversarial examples, generated by slightly crafting the legitimate samples, can mislead SVM classifier to give wrong classification results. Using the linear separable property of SVM in high-dimensional feature space, PSO is used to find the salient features, and then the average method is used to map back to the original input space to construct the adversarial example. This method makes full use of the easily finding salient features of linear models in the feature space, and the interpretable advantages of the original input space. Experimental results show that the proposed method can fool SVM classifier by using the adversarial example generated by less than 7 % small perturbation, thus proving that SVM has obvious security vulnerability.

Key words: Machine learning; Support Vector Machine(SVM); Exploring attacks; Salient perpetuation; Adversarial example

收稿日期：2018-08-06；改回日期：2019-01-28；网络出版：2019-02-15

*通信作者： 钱亚冠 QianYaGuan@zust.edu.cn

基金项目：浙江省自然科学基金(LY17F020011, LY18F020012)，浙江省公益技术应用研究项目(LGG19F030001)，国家自然科学基金(61772466, 61672337, 11771399)

Foundation Items: Zhejiang Natural Science Foundation (LY17F020011, LY18F020012), The Scientific Project of Zhejiang Provincial Science and Technology Department (LGG19F030001), The National Natural Science Foundation of China(61772466, 61672337, 11771399)

1　引言

随着以机器学习技术为核心的人工智能时代的到来，机器学习被广泛应用到安防、交通、医学、金融、气象、农业等领域，其中包括很多安全敏感的应用场景，例如自动驾驶中的视觉智能系统、视频监控中的目标识别系统、金融支付中的人脸识别系统等。尽管当前机器学习在这些领域取得了一些惊人的进展，人们对基于机器学习的智能系统内在脆弱性及其机理的理解仍处于初级阶段。研究人员已经开始注意到机器学习的安全问题并开始积极探索研究[1,2]。事实上，早在2006年，Barreno等人[1]指出由于机器学习自身的数据自适应性，机器学习模型可能在训练阶段或测试阶段遭到“毒性(poisoning)”攻击，也可能遭到以欺骗为目的的“探测(exploratory)”攻击。

随着深度神经网络在计算机视觉、语音识别等领域的成功应用，机器学习的安全性进一步引起了产学界的关注。Szegedy等人[3]在2014年首次指出深度神经网络存在决策“盲点”(blind spots)。即对一个图像加入微小的扰动，可使神经网络产生误判，却对人类视觉系统没有任何干扰作用。这种经过扰动篡改，使分类模型误判的数据被称为对抗样本(adversarial example)。神经网络自身的这种脆弱性引发了研究人员的极大兴趣，不少针对深度神经网络对抗样本的研究成果被发表[4–6]。

支持向量机(Support Vector Machine, SVM)作为一类重要的机器学习模型，也可以认为是特殊的单隐层神经网络[7]，在入侵检测、垃圾邮件分类、手写体识别等领域中得到广泛应用，其安全性也引起了研究人员的注意[8–10]。Biggio等人[8,9]通过随机扰动40%训练数据的标签，证明污染标签的毒性攻击足以降低SVM分类性能，并且又提出了在SVM训练集中插入恶意样本的毒性攻击方法，这些恶意样本是使用梯度上升方法寻找模型测试误差中的局部最大值相对应的输入。Mei等人[10]引入更通用的毒性攻击框架，对于SVM这种使用凸损失的模型，就可以根据Frobenius范数找到对训练集的最佳篡改。上述工作都是针对SVM的训练阶段展开毒性攻击及毒性攻击样本的生成。本文工作是在SVM的模型推断阶段实施探测性攻击。

Chen等人[11]提出基于梯度的攻击方法，针对高维SPAM(Substractive Pixel Adjancency Model)特征的SVM展开攻击，通过有限失真造成图像的决策错误，该攻击在直方图拉伸、自适应直方图均衡和中值滤波的检测中均有效。文献[12]利用Golland的辨别方向技术[13]启发的梯度下降方法来欺骗SVM。由于SVM中采用了核变换技巧，因此上述方法要求目标函数连续可微，在计算梯度时要先求取核梯度，计算代价会非常大。

本文提出SVM攻击显著性特征的概念，在容易处理线性模型的特征空间寻找攻击显著特征，再逆映射回输入空间完成原始样本的扰动，充分利用了特征空间上线性模型上易寻优的特点，同时又利用了原始输入空间篡改数据的可解释性优点。在实现过程中利用粒子群优化算法，不需要计算梯度，也就不要求目标函数连续可微，具有适应性强，容易实现的优点。实验表明，人造数据集中，2%扰动量下分类正确率降至50%以下；MNIST数据集的在5%扰动量下几乎完全分类错误；Yale人脸数据库在7%扰动量下分类正确率降至40%以下。

2　SVM分类器

SVM以其完备的理论基础和良好分类性能得到了广泛应用[14–16]。SVM模型的学习过程是一个寻找最优超平面的过程，最终形成的模型参数就是最优超平面的法向量。SVM训练的优化模型如式(1)

$$\begin{aligned}&\arg\min_{\boldsymbol{w}}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w},\\&\text{s.t.}\quad y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geq 1, i=1,2,\cdots,N\end{aligned}\tag{1}$$

其中，$\boldsymbol{x}_i$, y_i分别是训练数据和标记，训练数据为N个；$\boldsymbol{w}$, b是模型的学习参数，$\boldsymbol{w}$为d维向量，b为常数。考虑到实际情况通常是线性不可分，引入松弛变量ξ将式(1)转化为如式(2)的2次规划问题

$$\begin{aligned}&\arg\min_{\boldsymbol{w}}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+C\sum_{i=1}^{N}\xi_i,\\&\qquad i=1,2,\cdots,N\end{aligned}\tag{2}$$

其中，C为惩罚参数，ξ为惩罚项。当数据维度很大时，求解该问题时计算复杂度很高，为此将其转化为Lagrange对偶问题

$$\begin{aligned}&\arg\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}y_iy_j\alpha_i\alpha_j{\boldsymbol{x}_i}^{\mathrm{T}}\boldsymbol{x}_j-\sum_{i=1}^{N}\alpha_i,\\&\text{s.t.}\quad\sum_{i=1}^{N}y_i\alpha_i=0, 0\leq\alpha_i\leq C\end{aligned}\tag{3}$$

实际问题遇到的通常是线性不可分数据，因此需要将原始数据映射到高维空间$\boldsymbol{x}\to\phi(\boldsymbol{x})$获取线性超平面。由于高维空间的计算复杂度会很高[17]，可利用核函数$K(\boldsymbol{x},\boldsymbol{x}')=\phi(\boldsymbol{x})\phi(\boldsymbol{x}')$降低计算复杂度，在原始数据空间得到非线性决策边界。通过上述求解，从式(3)可以得到最终分类器的判决函数

$$F(\boldsymbol{x})=\mathrm{sign}\left(\sum_{i=1}^{N}\alpha_i y_i K(\boldsymbol{x}_i,\boldsymbol{x})+b\right) \tag{4}$$

3 问题的建立

通常针对学习系统的攻击方法可以分为毒性攻击和探测攻击[1]。其中，毒性攻击通过篡改训练数据影响模型的学习过程，使得学得的模型有利于对手达到攻击目标，而探测攻击则不同，它不影响最后模型的生成，而是在测试阶段躲避模型的识别从而实现其攻击目标。由于篡改训练数据的毒性攻击在现实中存在难度，相对来说探测攻击更容易实施，因此本文主要针对SVM的探测攻击展开研究。

由于本文的攻击模型接近“黑箱”攻击，唯一的假设是已知分类模型为基于多项式核的SVM。因此实际的攻击策略可以描述为如下步骤：(1)准备好与目标系统训练数据分布相近的数据；(2)向目标系统发送数据，获得返回的类标记；(3)通过上述步骤获得一个类似的训练数据集；(4)在此基础上在本地训练一个基于多项式核的SVM，其中多项式核的超参数通过经验确定；(5)在本地SVM上，利用优化方法生成攻击样本。如何生成能躲避或破坏SVM分类机制的攻击样本是攻击成功的关键，也是本文的重点研究部分。

生成对抗样本的基本策略是在原始样本的输入空间中加入扰动，在扰动最小的前提下，达到SVM误分类的目的。更正式的表达，假设$F:\boldsymbol{x}\to y$是SVM的判决函数，给定原始输入样本$\boldsymbol{x}$, y是它的类标记。在$\boldsymbol{x}$上加入扰动量$\delta_{\boldsymbol{x}}$后得到对抗样本$\tilde{\boldsymbol{x}}$。本文对抗样本生成策略建模为如式(5)的优化问题[3]

$$\arg\min_{\delta_{\boldsymbol{x}}}\|\delta_{\boldsymbol{x}}\|,\quad \text{s.t. } F(\boldsymbol{x}+\delta_{\boldsymbol{x}})\neq y \tag{5}$$

其中，$\boldsymbol{x}+\delta_{\boldsymbol{x}}=\tilde{\boldsymbol{x}}$。显然，对手通过在$\boldsymbol{x}$加入扰动$\delta_{\boldsymbol{x}}$后，让SVM分类器预测为$y$以外的标记。文献[4]提出寻找$\boldsymbol{x}$中最易干扰的特征是一种可行的方案，本文借鉴这种思路在SVM高维特征空间中寻找所谓的攻击显著性特征，容易利用线性模型易解的优点。

4 对抗样本生成

在输入特征空间中找到类似的攻击显著性特征，但SVM不会直接在输入特征空间构建线性模型，而是通过非线性映射到高维空间后构建线性决策超平面。因此，更好的方式就是在变换后的高维特征空间上寻找攻击显著性特征，然后逆映射回原始输入空间，构建对抗样本。本文假定SVM采用多项式核变换，$\phi(\boldsymbol{x}):\boldsymbol{x}\to\left(\sqrt{C_n^k a^{n-k}b^k}E^{(k)}\right)_k$，其中$a,b,n$为超参数，$k$为高维空间的第$k$项，$k=0,1,\cdots,n$，高维空间的维数为$1+d+\cdots+d^n$，$n\geq 2$，其中$E$通过式(6)的递推式得到

$$\left.\begin{aligned}
&E^{(0)}=1,\\
&E^{(1)}=\boldsymbol{x},\\
&E^{(2)}=(\boldsymbol{x}_i\odot\boldsymbol{x})=(\boldsymbol{x}_i\boldsymbol{x}_1,\boldsymbol{x}_i\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i\boldsymbol{x}_d)_i\\
&\quad=(\boldsymbol{x}_1\boldsymbol{x}_1,\cdots,\boldsymbol{x}_1\boldsymbol{x}_d,\cdots,\boldsymbol{x}_d\boldsymbol{x}_1,\cdots,\boldsymbol{x}_d\boldsymbol{x}_d),\\
&\quad i=1,2,\cdots,d\\
&E^{(k)}=(\boldsymbol{x}_i\odot E^{(k-1)}),k=3,4,\cdots,n
\end{aligned}\right\} \tag{6}$$

在变换空间训练SVM后得到的判别式为

$$F(\boldsymbol{x})=\mathrm{sign}(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})+b^*) \tag{7}$$

由于在变换空间中加入扰动会更直观也更好理解，式(5)中的对抗样本优化过程就变为

$$\left.\begin{aligned}
&\arg\min_{\delta_{\phi(\boldsymbol{x})}}\left\|\delta_{\phi(\boldsymbol{x})}\right\|\\
&\text{s.t.}\quad \mathrm{sign}(\boldsymbol{w}^{\mathrm{T}}(\phi(\boldsymbol{x})+\delta_{\phi(\boldsymbol{x})})+b^*)=\tilde{y}
\end{aligned}\right\} \tag{8}$$

式中，$\mathrm{sign}(s)$是指示函数，当$s>0$时$\mathrm{sign}(s)=+1$；当$s<0$时$\mathrm{sign}(s)=-1$。

考虑到求解上述优化问题是个NP难问题。文献[3]指出，线性模型的稳定性由最大权重系数决定，因此先采用粒子群优化策略获得特征空间的攻击显著性特征。由于本地的SVM模型的超参数已知，因此把攻击显著性特征空间经过多项式核的反变换到原始输入空间$\mathcal{X}$，然后利用均分操作得到原始输入空间的对抗样本。

粒子群优化(PSO)[18]是一种基于种群的启发式算法，它模拟社会行为，例如鸟类群集到有希望的位置，以期在多维空间中发现精确的目标。与进化算法一样，PSO使用个体(称为粒子)的群体(称为群)执行搜索，粒子从一个迭代更新到另一个迭代。为了找到最优解，每个粒子根据两个因素改变其搜索方向，即它自己的最佳先前位置(pbest)和所有其他成员的最佳位置(gbest)。Shi等人[19]称pbest为认知部分，gbest为社会部分。

寻找攻击显著特征的PSO算法如下述过程描述：首先，初始化粒子群，每个粒子在D维空间内具有随机位置，并且每个维度具有随机速度，第t次迭代的第i个粒子的位置和速度分别记为$\boldsymbol{x}_i^t=\{\boldsymbol{x}_{i1}^t,\boldsymbol{x}_{i2}^t,\cdots,\boldsymbol{x}_{iD}^t\}$, $\boldsymbol{v}_i^t=\{\boldsymbol{v}_{i1}^t,\boldsymbol{v}_{i2}^t,\cdots,\boldsymbol{v}_{iD}^t\}$。其次，评估每个粒子的适应度，这里的适应度函数使用$\mathrm{fit}(\boldsymbol{x})=\hat{\boldsymbol{w}}^{\mathrm{T}}\cdot\boldsymbol{z}$，其中$\hat{\boldsymbol{w}}$为显著性特征子集对应的$\boldsymbol{w}$的子集，适应度函数值越大，扰动效果越好，

第i个粒子在前t次迭代中的最佳适应度记为$\boldsymbol{p}_i^t=\{\boldsymbol{p}_{i1}^t,\boldsymbol{p}_{i2}^t,\cdots,\boldsymbol{p}_{iD}^t\}$。如果粒子的适应度优于最佳适应度，则更新最佳适应度。如果粒子的适应度比全局最佳适应度好，则适应度更新为全局最佳，前t次迭代的最佳适应度记为$\boldsymbol{p}_g^t=\{\boldsymbol{p}_{g1}^t,\boldsymbol{p}_{g2}^t,\cdots,\boldsymbol{p}_{gD}^t\}$。最后更新粒子的速度和位置，直到满足终止条件，其中，更新公式如式(9)和式(10)

$$\boldsymbol{v}_{id}^{t+1}=\boldsymbol{v}_{id}^t+c_1r_1(\boldsymbol{p}_{id}^t-\boldsymbol{x}_{id}^t)+c_2r_2(\boldsymbol{p}_{gd}^t-\boldsymbol{x}_{gd}^t) \tag{9}$$

$$\boldsymbol{x}_{id}^{t+1}=\boldsymbol{x}_{id}^t+\boldsymbol{v}_{id}^{t+1} \tag{10}$$

其中，$d=0,1,\cdots,D$，c_1为认知学习因子，c_2为社会学习因子，实验中均设置为2; r_1, r_2为[0, 1]间的随机数[20]。算法的伪代码如表1所示。

通过PSO方法得到特定于测试样本的扰动特征，即扰动对抗样本。但此时的扰动特征仍在多项式核变换的高维空间中，通过“均分”法将高维空间中的扰动分散到原始空间当中。由于对高维扰动子集作了索引，因此先确定子集中的每个特征在高维，即$1+d+\cdots+d^n$维空间中的具体位置，再根据E的定义，不难找出每个高维特征对应的原始空间中的特征；每个高维特征通常对应着多个原始空间特征，利用式(11)进行均分操作

$$\prod_{i\in I}(\boldsymbol{x}_i+\sigma)=\lambda\theta \tag{11}$$

其中，I为高维中需扰动的特征，λ为特征系数的倒数，σ为输入特征中的扰动。通过反变换，将扰动特征映射到输入空间，得到对抗样本的扰动。实现的伪代码如表2所示。

表 1　粒子群寻优(PSO)算法

输入：A //特征子集
输出：B //显著性特征

(1) $d=|A|, B=\phi$ //$A=(a^{(1)},a^{(2)},\cdots,a^{(d)})$
(2) FOR $i\leftarrow 1,2,\cdots,N$ DO
(3)　$\boldsymbol{s}_i\leftarrow\text{rand}(d),\boldsymbol{v}_i\leftarrow\text{rand}(d)$ //初始化N个粒子的位置和速度
(4)　$\boldsymbol{p}_i\leftarrow\boldsymbol{s}_i$ //$\boldsymbol{p}_i$为第i个粒子的当前最佳位置
(5) END FOR
(6) $\boldsymbol{p}_g\leftarrow\boldsymbol{p}_j$，其中$j\leftarrow\arg\max_i\text{fit}(\boldsymbol{p}_i), i=1,2,\cdots,N$ //$\boldsymbol{p}_g$为所有粒子的当前最佳位置
(7) FOR $k\leftarrow 1,2,\cdots,M$ DO //M为迭代次数
(8)　FOR $i\leftarrow 1,2,\cdots,N$ DO
(9)　$\boldsymbol{v}_{i+1}\leftarrow\boldsymbol{v}_i+c_1r_1(\boldsymbol{p}_i-\boldsymbol{s}_i)+c_2r_2(\boldsymbol{p}_g-\boldsymbol{s}_i)$
(10)　$\boldsymbol{s}_{i+1}\leftarrow\boldsymbol{s}_i+\boldsymbol{v}_{i+1}$
(11)　IF $\text{fit}(\boldsymbol{s}_{i+1})>\text{fit}(\boldsymbol{p}_{i+1})$ THEN
(12)　$\boldsymbol{p}_i\leftarrow\boldsymbol{s}_{i+1}$
(13)　END IF
(14) END FOR
(15) $\boldsymbol{p}_g\leftarrow\boldsymbol{p}_j$ 其中$j\leftarrow\arg\max_i\text{fit}(\boldsymbol{p}_i)$
(16) END FOR
(17) FOR $i\leftarrow 1,2,\cdots,d$ DO
(18) IF $\boldsymbol{p}_{g_i}>0.5$ THEN
(19)　$B\leftarrow B\cup\{a^{(i)}\}$ //$a^{(i)}$是$\boldsymbol{p}_{g_i}$对应的特征
(20)　END IF
(21) END FOR
(22) RETURN B

5　实验

5.1　数据集介绍

为了验证本文提出的对抗样本生成算法的可行性，本文采用MNIST[21]数据和Yale人脸数据库[22]2个数据集进行实验。

5.1.1　MNIST数据集

MNIST是一种广泛应用于机器学习测试手写体数据集，它共包含10个类别，分别为数字0至9。数据集共有70000个手写体数字图像，其中60000个为训练数据，10000个为测试数据。每个图像由28×28像素的大小构成，图1展示了几个手写体数字的图像示例。

5.1.2　Yale人脸数据集

这是一种常用于人脸识别的灰度人脸数据库，它共有165张人脸图像，包含15个人物，每个人物有11张人脸图像，由100×100像素的大小构成。图2展示了数据集中人脸图像示例。本文将每个人物的图像随机分为训练和测试两部分，其中7张(63.6%)

表 2　输入空间扰动算法

输入：A //$\boldsymbol{w}$从大到小排序后对应的特征
　　B //显著性特征
　　$\boldsymbol{X}_0$ //原始样本
输出：$\Delta\boldsymbol{X}$ //对抗样本的扰动

(1) $N=|B|,\Delta\boldsymbol{X}=\boldsymbol{0}$ //N为B的特征数，$\Delta\boldsymbol{X}$的大小与$\boldsymbol{X}_0$相同，且所有特征的初始值为0
(2) FOR $i\leftarrow 1,2,\cdots,N$ DO
(3)　$k\leftarrow\text{index}(b^{(i)})$ //k为$B=(b^{(1)},b^{(2)},\cdots,b^{(n)})$在特征空间的特征索引
(4)　$I\leftarrow\text{component}(k)$ // I为特征空间的第k个特征对应的“输入空间特征集”
(5)　$\sigma\leftarrow\delta(\theta,\lambda,I,\boldsymbol{X}_0)$//$\delta(\cdot)$由式(11)得到
(6)　FOR $j\leftarrow 1,2,\cdots,|I|$ DO
(7)　$\Delta\boldsymbol{X}(j)\leftarrow\Delta\boldsymbol{X}(j)+\sigma$
(8)　END FOR
(9) END FOR
(10) RETURN $\Delta\boldsymbol{X}$ //对抗样本的扰动

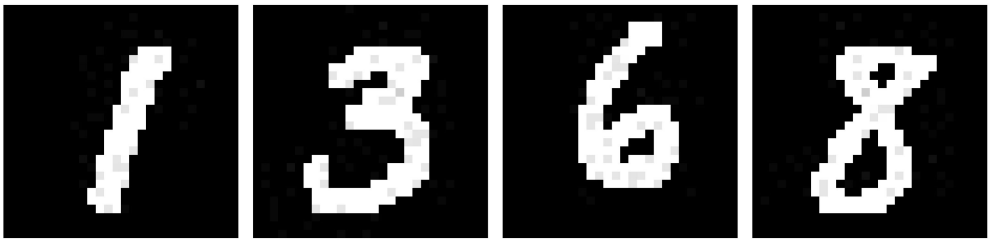

图 1 手写体数字图像示例

图 2 人脸图像示例

图像用于训练，4张用于测试，由于每个人物图像的训练和测试数据太少，对训练和测试数据分别通过旋转、翻转和添加噪声3种方式对图像样本进行扩充。扩充之后，每个人物的训练样本扩充至560个，测试样本至320个进行实验。

5.2 评估标准

对于MNIST数据集，通过给定不同扰动量计算SVM的攻击成功率，扰动量表示如式(12)

$$\delta = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\sum \boldsymbol{X}_1^{(i)} - \sum \boldsymbol{X}_0^{(i)}\right\|_1}{\left\|\sum \boldsymbol{X}_0^{(i)}\right\|_1} \tag{12}$$

其中，$\boldsymbol{X}_0$, $\boldsymbol{X}_1$分别为扰动前、后的测试样本，N为测试样本的个数。本文将设置4种不同的扰动量进行实验，其中人造数据取1%，2%，3%，4%，MNIST数据取1%, 3%, 5%和7%。

而对于较复杂的人脸数据库，采用改进的评估方法。根据文献[23]的“三庭五眼”理论将人脸图像分割为15块，本文中将长分割为33, 34, 33像素，宽均分为20像素，每块称为子图像。图2的示例图像分割后如图3所示。对同一人物的人脸作等量扰动的条件下，不同子图像呈现出的变化差别明显，用敏感度(SEN)来度量这一变化，计算公式如式(13)

$$\mathrm{SEN}_k = \frac{1}{n}\sum_{s=1}^{n}\frac{1}{pq-1}\sum_{i=1}^{p}\sum_{j=1}^{q}\left(\boldsymbol{v}_{ij}^{(s)} - \overline{\boldsymbol{v}^{(s)}}\right)^2 \tag{13}$$

其中，$k = 1, 2, \cdots, 15$为子图编号；p, q为子图的长、宽；$\boldsymbol{v}_{ij}$为子图在(i, j)位置上的像素值，$\overline{\boldsymbol{v}}$为子图的平均像素值，$n$为训练样本的个数。

为了防止不同子图间的敏感度差异过大，将M, m分别设定为最大、最小敏感度(可做无设定对照)，则敏感度可改写为

$$\overline{\mathrm{SEN}_k} = m + (M-m)\cdot\frac{\mathrm{SEN}_k - \bar{m}}{\bar{M} - \bar{m}} \tag{14}$$

其中，$\bar{m} = \min_k\{\mathrm{SEN}_k\}$, $\bar{M} = \max_k\{\mathrm{SEN}_k\}$。将敏感度作为权重，即加权扰动量来调整式(12)

$$\bar{\delta} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\sum \overline{\mathrm{SEN}_k}\boldsymbol{X}_1^{(i)} - \sum \overline{\mathrm{SEN}_k}\boldsymbol{X}_0^{(i)}\right\|_1}{\left\|\sum \overline{\mathrm{SEN}_k}\boldsymbol{X}_0^{(i)}\right\|_1} \tag{15}$$

其中，$\bar{\boldsymbol{X}}_p^i = \sum \boldsymbol{X}_p^{(k)}, p = 0, 1,\ k = 1, 2, \cdots, 15$。扰动量的设置与MNIST数据集一致，即1%, 3%, 5%和7%。通常扰动量越大，人眼越容易发觉，通过对不同扰动量下的图像对比，能够粗略得到人眼识别人脸图像的扰动量阈值为7%。

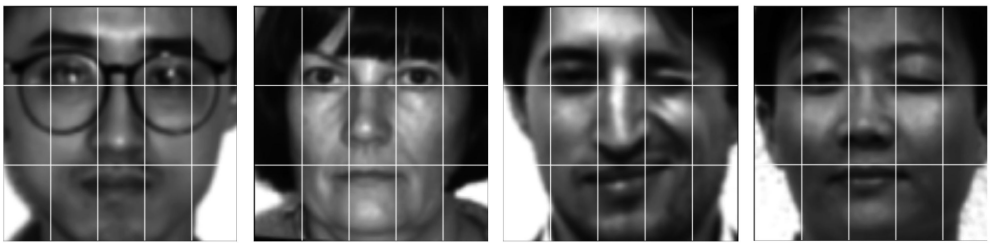

图 3 “三庭五眼”的人脸分割示例

5.3 攻击效果

5.3.1 MNIST

在训练模型阶段，使用网格寻优方法设置多项式核SVM的超参数。训练好模型后，对测试集中各个类别的手写体的分类正确率单独测试，结果如表3所示。总体分类正确率达95.77%；最低的类别为数字5，但也达到了92.71%，最高为数字1，达到98.94%，超过总体分类正确率的包括数字0, 1, 4和6。

表 3　测试集中各个手写体的分类准确率(%)

手写体数字	0	1	2	3	4	5	6	7	8	9
准确率	98.88	98.94	95.16	95.74	96.13	92.71	97.18	94.65	93.94	93.76

在不同扰动量下对各个类别的测试样本作扰动攻击，结果如表4所示。1%的少量扰动量对大多数手写体数字的攻击效果不显著。较明显的为数字3，分类正确率下降13.98%，最不显著的为数字8，仅仅下降0.52%。当扰动量提高到3%和5%时，对大多数数字可产生显著攻击效果，分类正确率平均下降接近50%。同时，手写体的各个数字对扰动量的敏感程度不同，如数字1在3%的扰动量下分类正确率降至30%左右，5%的扰动量下几乎完全分类错误。而一些数字在扰动量增至7%时仍有较高的分类正确率，如数字2接近60%，这可能和数字2与其他类别的相似度不高引起的。数字2在扰动前图4(a)和不同扰动量图4(b)—图4(e)下的图像如图4所示。

5.3.2 Yale人脸数据库

实验中，将每个人物的560张图像作为训练集，320张作为测试集。为了消除扰动图像时随机性带来的影响，对每一个测试样本均作多次实验，并以平均值作为最后的结果。

表 4　不同扰动量下各类手写体数字的平均分类正确率(%)

手写体数字	扰动前	1%扰动	3%扰动	5%扰动	7%扰动
0	98.88	95.32	75.37	37.44	10.17
1	98.94	96.48	31.93	13.57	1.21
2	95.16	84.54	72.14	64.93	58.65
3	95.74	81.76	67.89	50.22	30.74
4	96.13	92.44	42.98	8.76	0.39
5	92.71	89.38	55.73	18.37	5.65
6	97.18	94.63	70.64	30.58	12.33
7	94.65	91.71	69.87	32.43	17.47
8	94.65	94.13	78.21	35.38	13.58
9	93.94	90.85	52.73	27.64	6.53

在给定扰动量下，不同的人脸图像对扰动区域(扰动子图)的选择不同，图5展示了原始图像、扰动及扰动后生成的对抗样本。图5第1行为原图，第2行为扰动噪声，第3行为对抗样本。图6展示了不同比例下的扰动量，从左至右分别为原图，1%, 3%, 5%和7%的扰动分布。

对每个人物的人脸测试图像进行了50至60次不等的实验，取平均后得到结果如表5所示。当扰动量取1%时，人物7, 11和13的分类正确率低于85%。当扰动量取3%时，人物7的分类正确率略高于40%。所有人物在7%的扰动量下的分类正确率均不足40%，其中人物13的分类正确率已低至6.37%。很明显，随着扰动量的提高，分类正确率呈现下降趋势，当扰动量从1%增加至3%时，人物7分类准确率下降最显著，达到42.61%；当扰动量从3%增加至5%时，人物11准确率下降37.1%；当扰动量从5%增加至7%时，分类准确率下降最显著的是人物14，达到41.45%。不同扰动量对人物7, 11, 13和14较为敏感，对分类正确率的影响较大，而人物1, 12对扰动的稳定性较好。这可能由于人物7, 13和14的人脸图像中包含部分配饰，如眼镜。人物11是所有对象中唯一的女性，也可能是其中的重要因素。而人物1和12较其他对象而言，人脸特征不够突出反而成为抗扰动的稳定因素。

6　结束语

SVM是一类重要的机器学习模型，被广泛应用于各种安全敏感领域，它的安全性必须引起足够

(a)扰动前　(b) 1%扰动　(c) 3%扰动　(d) 5%扰动　(e) 7%扰动

图 4 不同扰动程度的图像示例

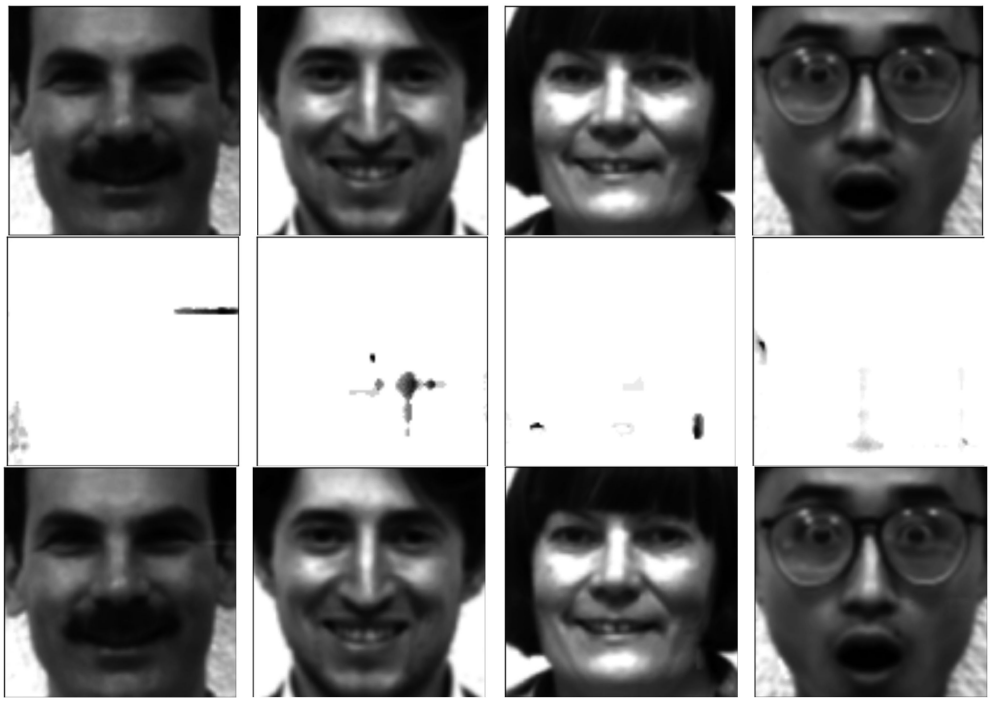

图 5 人脸扰动前后的图像示例

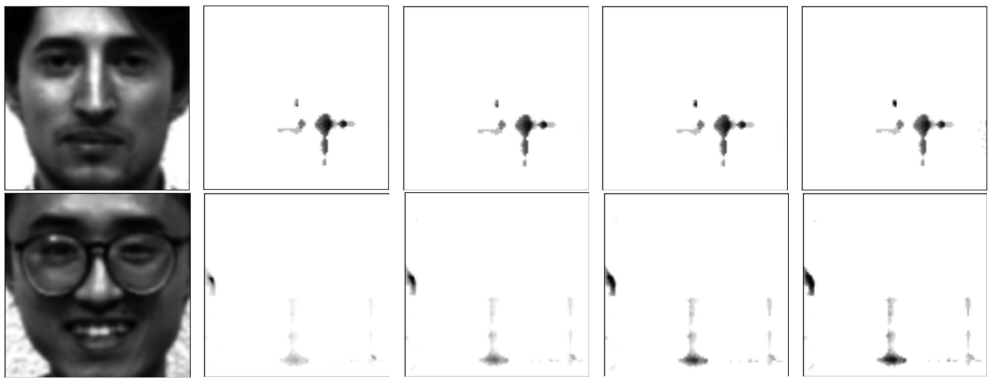

图 6 不同扰动量下的对象示例

表 5 不同扰动比例下各对象的平均分类正确率(%)

人脸序号	1%扰动	3%扰动	5%扰动	7%扰动
1	95.12	90.02	68.82	38.63
2	87.68	71.13	54.98	29.22
3	91.19	81.57	58.13	29.16
4	89.43	75.27	52.29	21.09
5	90.78	79.27	43.55	26.87
6	87.91	71.62	60.14	21.33
7	83.26	41.12	15.67	8.31
8	92.43	70.22	47.93	29.83
9	91.33	75.71	46.62	28.11
10	94.66	81.73	57.45	30.13
11	82.63	68.20	30.79	10.32
12	98.78	81.17	66.05	37.16
13	72.65	57.27	33.48	6.37
14	85.17	63.33	49.78	7.91
15	97.5	89.85	70.21	29.84

的重视。针对可能的探测攻击，本文提出了基于粒子群优化的对抗样本生成算法，并通过两个实际数据集对其攻击能力进行了验证。实验结果表明，本文提出的算法从简单的人造数据，到复杂的人脸数据均具有良好的攻击效果，从另一方面也揭示出SVM的安全脆弱性。因此，下一阶段以此为基础，积极探索SVM的安全学习机制，提高其抗攻击能力。

参 考 文 献

[1] BARRENO M, NELSON B, SEARS R, *et al*. Can machine learning be secure?[C]. Proceedings of 2006 ACM Symposium on Information, Computer and Communications Security, Taipei, China, 2006: 16–25. doi: 10.1145/1128817.1128824.

[2] LI Pan, ZHAO Wentao, LIU Qiang, *et al*. Security issues

and their countermeasuring techniques of machine learning: A survey[J]. *Journal of Frontiers of Computer Science & Technology*, 2018, 12(2): 171–184.

[3] SZEGEDY C, ZAREMBA W, SUTSKEVER I, *et al*. Intriguing properties of neural networks[EB/OL]. http://arxiv.org/abs/1312.6199v4, 2014.

[4] PAPERNOT N, MCDANIEL P, JHA S, *et al*. The limitations of deep learning in adversarial settings[C]. Proceedings of 2016 IEEE European Symposium on Security and Privacy, Saarbrucken, Germany, 2016: 372–387. doi: 10.1109/EuroSP.2016.36.

[5] PAPERNOT N, MCDANIEL P, GOODFELLOW I, *et al*. Practical black-box attacks against machine learning[EB/OL]. http://arxiv.org/abs/1602.02697v4, 2017.

[6] AKHTAR N and MIAN A. Threat of adversarial attacks on deep learning in computer vision: A survey[J]. *IEEE Access*, 2018, 6: 14410–14430. doi: 10.1109/ACCESS.2018.2807385.

[7] CORTES C and VAPNIK V. Support-vector networks[J]. *Machine Learning*, 1995, 20(3): 273–297. doi: 10.1007/BF00994018.

[8] BIGGIO B, NELSON B, and LASKOV P. Support vector machines under adversarial label noise[C]. Proceedings of the 3rd Asian Conference on Machine Learning, Taoyuan, China, 2011, 20: 97–112.

[9] BIGGIO B, NELSON B, and LASKOV P. Poisoning attacks against support vector machines[EB/OL]. http://arxiv.org/abs/1206.6389v3, 2013.

[10] MEI Shike and ZHU Xiaojin. Using machine teaching to identify optimal training-set attacks on machine learners[C]. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, USA, 2015: 2871–2877.

[11] CHEN Zhipeng, TONDI B, LI Xiaolong, *et al*. A gradient-based pixel-domain attack against SVM detection of global image manipulations[C]. Proceedings of 2017 IEEE Workshop on Information Forensics and Security, Rennes, France, 2017: 1–6. doi: 10.1109/WIFS.2017.8267668.

[12] BIGGIO B, CORONA I, MAIORCA D, *et al*. Evasion attacks against machine learning at test time[EB/OL]. http://arxiv.org/abs/1708.06131, 2013.

[13] GOLLAND P. Discriminative direction for kernel classifiers[C]. Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, Vancouver, British Columbia, Canada, 2001: 745–752.

[14] AMRAEE S, VAFAEI A, JAMSHIDI K, *et al*. Abnormal event detection in crowded scenes using one-class SVM[J]. *Signal, Image and Video Processing*, 2018, 12(6): 1115–1123. doi: 10.1007/s11760-018-1267-z.

[15] BENMAHAMED Y, TEGUAR M, and BOUBAKEUR A. Application of SVM and KNN to Duval pentagon 1 for transformer oil diagnosis[J]. *IEEE Transactions on Dielectrics and Electrical Insulation*, 2017, 24(6): 3443–3451. doi: 10.1109/TDEI.2017.006841.

[16] SCHNALL A and HECKMANN M. Feature-space SVM adaptation for speaker adapted word prominence detection[J]. *Computer Speech & Language*, 2019, 53: 198–216. doi: 10.1016/j.csl.2018.06.001.

[17] ZHAO Rui and MAO Kezhi. Semi-random projection for dimensionality reduction and extreme learning machine in high-dimensional space[J]. *IEEE Computational Intelligence Magazine*, 2015, 10(3): 30–41. doi: 10.1109/MCI.2015.2437316.

[18] EBERHART R and KENNEDY J. A new optimizer using particle swarm theory[C]. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, Nagoya, Japan, 2002: 39–43. doi: 10.1109/MHS.1995.494215.

[19] SHI Y and EBERHART R. A modified particle swarm optimizer[C]. Proceeding of 1998 IEEE International Conference on Evolutionary Computation, World Congress on Computational Intelligence, Anchorage, USA, 1998: 69–73. doi: 10.1109/ICEC.1998.699146.

[20] LIN S W, YING K C, CHEN S C, *et al*. Particle swarm optimization for parameter determination and feature selection of support vector machines[J]. *Expert Systems with Applications*, 2008, 35(4): 1817–1824. doi: 10.1016/j.eswa.2007.08.088.

[21] LECUN Y, CORTES C, and BURGES C J C. The MNIST database of handwritten digits[EB/OL]. http://yann.lecun.com/exdb/mnist/, 2010.

[22] YALE. The Yale face database[OL]. http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html, 1997.

[23] 何光辉, 唐远炎, 房斌, 等. 图像分割方法在人脸识别中的应用[J]. 计算机工程与应用, 2010, 46(28): 196–198. doi: 10.3778/j.issn.1002-8331.2010.28.055.
HE Guanghui, TANG Yuanyan, FANG Bin, *et al*. Image partition method in face recognition[J]. *Computer Engineering and Applications*, 2010, 46(28): 196–198. doi: 10.3778/j.issn.1002-8331.2010.28.055.

钱亚冠：男，1976年生，副教授，研究方向为机器学习安全、计算机视觉.

卢红波：男，1993年生，硕士生，研究方向为机器学习安全.

纪守领：男，1986年生，研究员，主要研究方向为人工智能安全、数据驱动安全与隐私保护.

周武杰：男，1983年生，副教授，主要研究方向为机器视觉.

吴淑慧：女，1975年生，讲师，研究领域为深度神经网络.

云本胜：男，1980年生，讲师，研究领域为机器学习与数据挖掘.

陶祥兴：男，1966年生，教授，主要研究领域为信号处理与金融数据分析.

雷景生：男，1967年生，教授，主要研究领域为机器学习与大数据处理.