



一种基于贪心算法的 SVM 扰动攻击方法

钱亚冠¹, 关晓惠², 吴淑慧¹, 云本胜¹, 任东晓¹

(1. 浙江科技学院, 浙江 杭州 310023;

2. 浙江水利水电学院, 浙江 杭州 310018)

摘要: 随着对机器学习安全问题的关注度不断提高, 提出一种针对 SVM (support vector machine, 支持向量机) 的攻击样本生成方法。这种攻击发生在测试阶段, 通过篡改实例数据, 达到欺骗 SVM 分类模型的目的, 具有很大的隐蔽性。采用贪心策略在核空间中搜索显著性特征子集; 然后将核空间中的扰动映射回输入空间, 获得攻击样本。该方法通过不超过 7% 的小扰动量使测试样本错误地分类。对 2 个数据集进行实验, 攻击均能取得成功。在人造数据集中, 2% 的扰动量下可使 SVM 分类器的错误率在 50% 以上; 在 MNIST 数据集中, 5% 的扰动量可使 SVM 分类器错误率接近 100%。

关键词: 机器学习; 支持向量机; 贪心算法; 显著性特征; 攻击样本

中图分类号: TP309.2

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019014

A novel perturbation attack on SVM by greedy algorithm

QIAN Yaguan¹, GUAN Xiaohui², WU Shuhui¹, YUN Bensheng¹, REN Dongxiao¹

1. Zhejiang University of Science and Technology, Hangzhou 310023, China

2. Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China

Abstract: With the increasing concern of machine learning security issues, an adversarial sample generation method for SVM was proposed. This attack occurred in the testing stage by manipulating with the sample to fool the SVM classification model. Greedy strategy was used to search for salient feature subsets in kernel space and then the perturbation in the kernel space was projected back into the input space to obtain attack samples. This method made the test samples misclassified by less than 7% perturbation. Experiments are carried out on two data sets, and both of them are successful. In the artificial data set, the classification error rate is above 50% under 2% perturbation. In MNIST data set, the classification error rate is close to 100% under 5% perturbation.

Key words: machine learning, support vector machine, greedy algorithm, salient feature, adversarial sample

收稿日期: 2018-07-24; 修回日期: 2018-10-16

基金项目: 浙江省自然科学基金资助项目 (No.LY17F020011); 浙江省公益技术应用研究项目 (No.LGG19F030001); 国家自然科学基金资助项目 (No.61572163)

Foundation Items: The Natural Science Foundation of Zhejiang Province of China (No.LY17F020011), The Scientific Project of Zhejiang Provincial Science and Technology Department (No.LGG19F030001), The National Natural Science Foundation of China (No.61572163)



1 引言

2014 年, Szegedy 等^[1]发表了一篇非常有影响力的论文, 宣称深度神经网络存在决策“盲点(blind spot)”。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。经过扰动篡改使分类模型误判的数据被称为攻击样本(adversarial sample)。这种神经网络自身的脆弱性引发了研究人员的极大兴趣, 针对深度神经网络攻击样本的研究成果大量出现^[2-4]。神经网络的这种脆弱性向当前所有的机器学习方法自身的安全性提出了挑战, 一旦被应用到无人驾驶、身份识别等安全领域, 其自身可能会遭到攻击。事实上, 早在 2006 年, 美国加州大学伯克利分校的 Barreno 等^[5]提出了“机器学习安全吗?”, 指出由于机器学习自身的数据自适应性, 可能在训练阶段或测试阶段遭到“毒性”攻击, 也可能遭到以欺骗为目的的“探测”攻击。

支持向量机(support vector machine, SVM)作为一类重要的机器学习模型, 可以认为是特殊的单隐层神经网络^[6], 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^[7-8], 其安全性也引起了研究人员的注意^[9-11]。Biggio 等^[12]在这一领域做了许多出色的工作, 包括在知道模型架构及参数的情况下进行 SVM 攻击和对学习模型的训练数据进行“毒性”攻击^[13]等。此外, Lin 等^[14]研究了“白箱”条件下对 SVM 的“毒性”攻击, 大大降低了 SVM 模型的精度。Liu 等^[15]改进了 Smith-Waterman 方法, 使用 SVM 分类器建立了新的攻击特征生成器模型, 降低了系统警报率。Chen 等^[16]提出针对高维垃圾邮件特征的基于梯度的 SVM 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。

本文与上述工作不同, 主要研究测试分类阶

段的欺骗攻击^[2]。在深度学习分类任务中, 对 MNIST 数据集中的样本加入不足 4.5%的微小扰动就能显著地改变分类结果, 误分类比例超过 97%, 该类以欺骗为目的的攻击称为扰动攻击, 经过扰动的数据样本称为攻击样本。受参考文献^[2]的启发, 笔者针对 SVM 展开扰动攻击方面的研究, 提出 SVM 上同样存在输入输出间的攻击显著性特征的假设, 并由此采用贪心算法构建攻击样本。实验结果证明这种攻击显著性的特征不仅存在于深度神经网络中, 同样存在于支持向量机中, 这给笔者下一步探索基于机器学习的智能系统的安全性提供了进一步的研究空间。

2 SVM 分类器

SVM 以其完备的理论基础和良好的分类性能得到了广泛应用^[17-18]。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即最优超平面的法向量。SVM 训练的优化模型如下:

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

其中, $\mathbf{x}_i$ 、 y_i 分别是训练数据和标记, 训练数据为 N 个; $\mathbf{w}$ 、 b 是模型的学习参数, $\mathbf{w}$ 为 d 维向量, b 为常数。考虑到实际情况通常是线性不可分的, 引入松弛变量 ξ 将式 (1) 转化为二次规划问题, 具体如下:

$$\begin{aligned} \arg \min_{\mathbf{w}} & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i \\ \text{s.t. } & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \dots, N \end{aligned} \quad (2)$$

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, 为此将其转化为拉格朗日对偶问题:

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

实际中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 $\mathbf{x} \rightarrow \phi(\mathbf{x})$ 以获取线性超平面。由于高维空间的计算复杂度很高^[20], 可利用核函数 $K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})\phi(\mathbf{x}')$ 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, 可以从式 (3) 得到最终分类器的判决函数:

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

3 针对 SVM 的攻击策略

通常针对学习系统的攻击方法可以分为“毒性”攻击和扰动攻击^[5]。其中,“毒性”攻击通过篡改训练数据影响模型的学习过程,使学习所得的模型有利于对手达到攻击目标;而扰动攻击则不同,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达到其攻击目标。由于篡改训练数据的“毒性”攻击在现实中存在难度,相对来说扰动攻击更容易实施,因此本文主要针对 SVM 的扰动攻击展开研究。

3.1 攻击模型

根据参考文献[2],从攻击目标和攻击能力两方面面对扰动攻击进行建模。攻击目标从易到难依次可分为 4 类:降低分类器输出的置信度;把输出误分到其他类,如把手写体“7”误分到任意其他数字类;误分到指定的目标类,如把某个手写体数字强制误分到“5”;特定的输入被强制误分到指定的目标类,如必须把手写体“7”强制分类到“5”。

针对学习系统的攻击能力与对手能掌握目标系统信息的多少有关。通常信息掌握得越多,对手实施攻击的能力就会越强。根据对手对分类系统的了解程度,由少到多可分为 4 类:(1)对手不知道训练数据和模型参数等信息,但对手可以向分类系统发送测试数据,并会返回输出类别,在这种情况下实施的攻击称为“黑箱”攻击;(2)对手可以收集到与目标模型训练数据同

分布的数据集,但不知道模型及其参数;(3)对手知道目标模型,但不知道具体的模型参数;(4)对手知道模型的所有信息,包括训练数据和模型参数,这种情况下实施的攻击称为“白箱”攻击。显然,对手的这 4 种能力层次呈包含关系:(1) $\subset$ (2) $\subset$ (3) $\subset$ (4)。

本文的攻击模型假设对手的能力处于层次(3),即对手知道分类器模型,但无法直接获取和操纵训练集,不能影响分类模型的生成,无法直接获取模型的参数。攻击目标是把输入样本误分类。

3.2 攻击策略

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流程如图 1 所示,具体如下:

- (1) 准备好与目标系统训练数据分布相近的数据,并向目标系统发送数据;
- (2) 获得远程系统响应返回的类标签;
- (3) 通过上述步骤获得一个类似的训练数据集,并在此基础上在本地训练一个基于多项式核的 SVM;
- (4) 获得本地 SVM 模型的超参数,生成攻击样本;
- (5) 向远程目标 SVM 发送攻击样本,实施攻击。

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。笔者在实验室用两台计算机分别模拟本地 SVM 和远程目标系统,以实现图 1 的攻击步骤。

生成攻击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(保证不被人工察觉)的前提下,达到 SVM 误分类的目的。更正后的表达为,假设 $F: \mathbf{x} \rightarrow y$ 是 SVM 的判决函数,给定原始输入样本 $\mathbf{x}$, y 是其类标记。在 $\mathbf{x}$ 上加入扰动量 $\delta_{\mathbf{x}}$ 后得到攻击样本 $\tilde{\mathbf{x}}$ 。把上述攻击样本生成策略建模为如下优化问题^[1]:

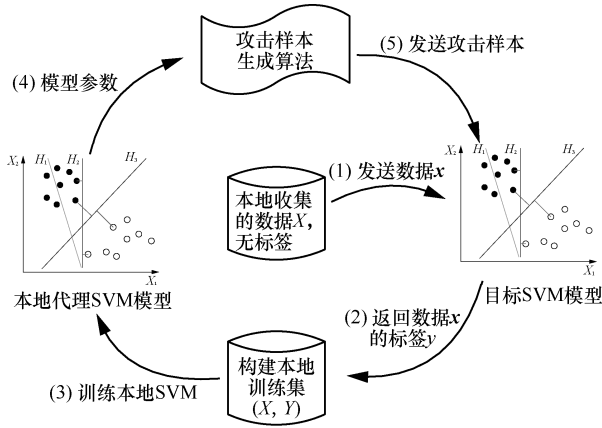


图1 实际的攻击策略流程

$$\begin{aligned} \arg \min_{\delta_x} \|\delta_x\| \\ \text{s.t. } F(\mathbf{x} + \delta_x) \neq y \end{aligned} \quad (5)$$

其中, $\mathbf{x} + \delta_x = \tilde{\mathbf{x}}$ 。显然, 对手通过在 $\mathbf{x}$ 加入扰动 δ_x 后, 让 SVM 分类器预测为 y 以外的标记。参考文献[2]提出寻找 $\mathbf{x}$ 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 分类器中寻找所谓的攻击显著性特征。

4 基于贪心算法的攻击样本生成方法

参考文献[2]通过计算 DNN 的前向导数, 构建所谓的攻击显著性图 (adversarial saliency map) 识别与对手攻击目标显著相关的输入特征。把干扰加到这些特征上可以很快导致期望的攻击性输出, 例如误分类。受该文的启发, 笔者希望在 SVM 的输入特征空间中找到类似的攻击显著性图。参考文献[1]指出, 线性模型的稳定性是由最大权重系数决定的。实际的 SVM 通常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。因此, 最直接的方式是在变换后的高维特征空间上构建攻击显著性图。形式化地描述这个问题, 本文采用多项式核变换, $\phi(\mathbf{x}): \mathbf{x} \rightarrow \left(\sqrt{C_n^k} a^{n-k} b^k E^{(k)} \right)_k$, 其中, a 、 b 、 n 为超参数, k 为高维空间的第 k 项, $k=0, 1, \dots, n$, 高维空间的维数为 $1+d+\dots+d^n$, $n \geq 2$ 。在变换空间训练 SVM 后得到的判别式为:

$$F(\mathbf{x}) = \text{sign}(W^T \phi(\mathbf{x}) + b^*) \quad (6)$$

由于在变换空间中加入扰动会更直观也更好理解, 式 (5) 中的攻击样本优化过程变为:

$$\begin{aligned} \arg \min_{\delta_{\phi(\mathbf{x})}} \|\delta_{\phi(\mathbf{x})}\| \\ \text{s.t. } \text{sign}(W^T (\phi(\mathbf{x}) + \delta_{\phi(\mathbf{x})}) + b^*) = \tilde{y} \end{aligned} \quad (7)$$

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

基于多项式核的 SVM 利用核函数在原始的输入空间获得非线性决策边界, 在高维特征空间上构建线性超平面。线性超平面更有利于寻找攻击显著性特征, 但由于高维空间在计算上难度很大, 为此仍然需要借助于核技巧在原始输入空间中实现求解。考虑到求解的问题是 NP 难问题, 笔者采用贪心策略^[21]获得总体样本的显著性特征子集。由于已知本地的 SVM 模型的超参数, 因此由贪心算法得到的扰动攻击样本经过多项式核反变换到原始空间 X , 即攻击样本。

首先设定扰动因子 θ 来限制每个特征允许的最大改变量, 设定扰动比例 ε 来限制允许扰动的特征数占总特征数的最大比例。采用贪心策略, 首先将最大权重对应的特征选择为可扰动特征, 逐步增加扰动量, 在达到扰动因子 θ 限制前, 如果攻击成功则停止; 否则不断挑选剩余最大权重对应的特征, 重复上述加扰过程, 直到完成攻击或达到最大扰动比例上限。

形式化地描述上述算法过程: 给定样例 $\mathbf{x} \in R^n$, 用 $\phi(\mathbf{x})$ 变换到高维特征空间 $(z_1, \dots, z_m)$, $z_i \in R$ 是特征空间中第 i 个特征的值, 变换后的特征空间维数为 m , 通常 $n \ll m$ 。判决式 (6) 可写为:

$$F(\mathbf{z}) = \text{sign}(w^{(1)} z^{(1)} + \dots + w^{(m)} z^{(m)} + b) \quad (8)$$

假定 $\mathbf{x}$ 的类标签 $y = -1$, 攻击的目的是让

SVM 把扰动后的 $\mathbf{x}$ 误分类为 $y = +1$ 。对式 (6) 中的 w 从大到小排序为 $(w^{(1)}, \dots, w^{(m)})$ ，并将其对应特征和特征值记为 $(a^{(1)}, \dots, a^{(m)})$ 和 $(z^{(1)}, \dots, z^{(m)})$ 。攻击显著性特征子集 A 通过如下过程得到：初始化子集 $A = \emptyset$ 。首先对特征值 $z^{(1)}$ 逐次加扰 $z^{(1)} \leftarrow z^{(1)} + \gamma z^{(1)}$ ，其中 γ 为单次加扰幅度， $A = \{a^{(1)}\}$ 。把加扰后的样例 z' 代入判别式 (6)，若 $F(z') = +1$ ，或 $\sum \gamma$ 达到上限 θ ，则停止扰动；否则在原来扰动结果 z' 的基础上，继续对第二个特征值加扰 $z^{(2)} \leftarrow z^{(2)} + \gamma z^{(2)}$ ， $A = \{a^{(1)}, a^{(2)}\}$ 。重新计算 $F(z'')$ ，重复上述逐次加扰过程，直到满足 $F(z'') = +1$ 或者扰动比例 $\frac{|A|}{m} < \varepsilon$ 。显然， ε 的选择非常重要，取太大的值容易攻击成功，但也容易发现原始数据被篡改；取太小的值，尽管扰动痕迹不易发现，但也难以使分类器判决失效。因此， ε 作为一个超参数，需要经过多次实验尝试才能确定。实现的伪代码如下。

算法 1 greedy_search //显著性特征子集搜索算法

输入： $\bar{A}$ ， w 从大到小排序后对应的特征；
 Y ，测试样本的标记；

Z ，特征空间中的原始样本。

输出： A ，显著性特征子集。

$A \leftarrow \emptyset, Z \leftarrow (z^{(1)}, z^{(2)}, \dots, z^{(m)})$ // Z 为 w 从大

到小排序后对应的特征值

for $i \leftarrow 1, 2, \dots, m$ **do**

$A \leftarrow A \cup \{a^{(i)}\}$ //从 $\bar{A}$ 中取出特征 $a^{(i)}$

添加到 A ， $\bar{A} = (a^{(1)}, a^{(2)}, \dots, a^{(m)})$

$\Delta z^{(i)} \leftarrow \gamma$ // γ 为单次扰

动量

while $\Delta z^{(i)} < \theta$ **do** // $\Delta z^{(i)}$ 第 i 个特征的累积改变量， θ 为扰动上限

$z^{(i)} \leftarrow z^{(i)} + \gamma$

if $F(Z) \neq Y$ **then break** // Z_0 为当前

扰动样本， F 为判别函数

end if

$z^{(i)} \leftarrow z^{(i)} + \gamma$

end while

if $\frac{|A|}{m} > \varepsilon$ **then break**

end if

end for

return A // 返回显著性特征子集

通过贪心算法得到特定于测试样本的显著性特征。但此时的扰动特征仍在多项式核变换的高维空间中，笔者通过“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，因此本文先确定子集中的每个特征在高维（即 $1 + d + \dots + d^n$ 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，本文利用式 (9) 进行均分操作：

$$\prod_{i \in I} (x_i + \sigma) = \lambda \theta \quad (9)$$

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，得到输入空间的攻击样本。

5 实验与分析

5.1 数据集介绍

为了验证本文提出的攻击样本生成算法的可行性，笔者采用人造随机数据和 MNIST^[22] 手写体这两个数据集进行实验。

人造随机数据集由 MATLAB 的随机数方法生成二维数据，数据范围分别为 $[-3, 4]$ 和 $[-1, 2.5]$ ，共生成正例、负例各 1 600 个；将正例和负例随机等分为两部分，分别作为训练集和测试集使用；训练集在二维坐标下的分布如图 2 所示，其中“o”和“x”分别表示正例和负例。

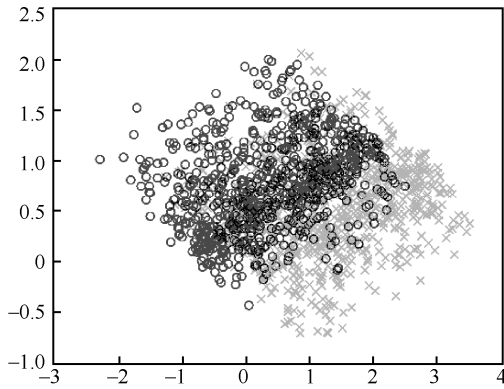


图2 人造随机数据集的训练集

MNIST 数据集是一种受到广泛欢迎的手写体数据集,共包含 10 个类别,分别为类别 0~类别 9;数据集共含有 70 000 个图像,其中 60 000 个为训练图像,10 000 个为测试图像,每个图像由 28×28 像素构成,图 3 展示了几个手写体类别的图像示例。

5.2 评估方法

对于人造数据集与 MNIST 数据集,本文通过给定不同扰动量计算 SVM 的攻击成功率,扰动量表示如下:

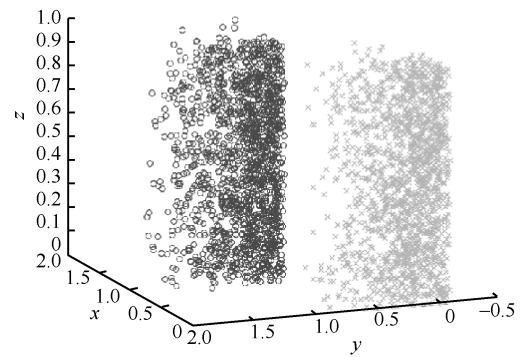
$$\delta = \frac{1}{N} \sum_{i=1}^N \frac{\|\sum x_1^{(i)} - \sum x_0^{(i)}\|_1}{\|\sum x_0^{(i)}\|_1} \quad (10)$$

其中, x_0 、 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,前者为 1%、2%、3%、4%,后者为 1%、3%、5%和 7%。

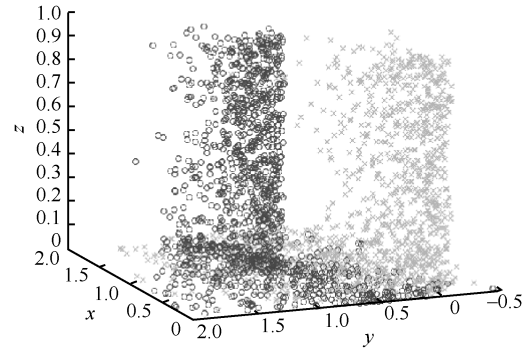
5.3 攻击效果

5.3.1 人造随机数据集

显然,图 2 中的训练集无法在二维空间下线性分类,通过变换将数据集映射到三维特征空间,如图 4 (a) 所示,本文找到了线性超平面将两种形状进行较好的线性分类。本文通过对同分布下的测试集做扰动进行攻击,其中测试集在扰动量为 1%下的三维特征空间结果如图 4 (b) 所示,不同干扰量下的分类成功率见表 1。



(a) 原始数据



(b) 1%扰动

图4 三维特征空间下的训练集和 1%扰动量下的测试集

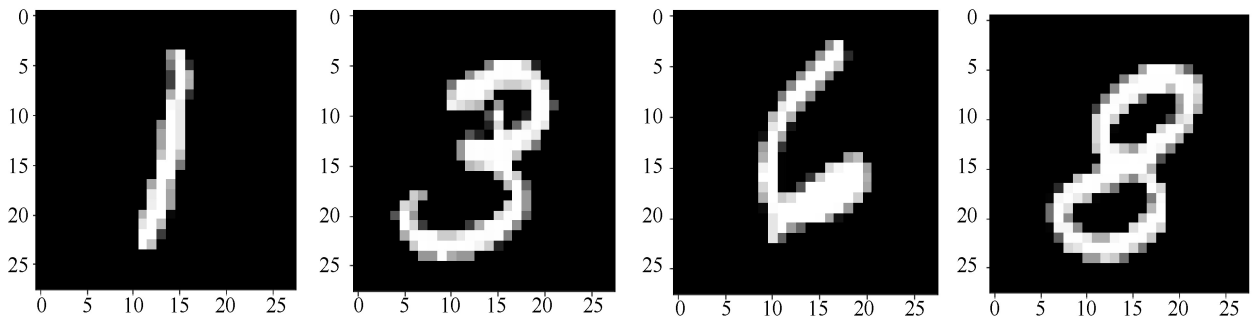


图3 手写体图像示例

表1 不同扰动量下的分类成功率

原测试集	1%	2%	3%	4%
99.5%	87%	50.5%	36%	29.5%

由表1的结果可知,测试样本在正常条件下的分类错误率仅为0.5%,而扰动条件下的分类成功率有不同程度的下降。扰动量为1%时,有13%的测试样本被错误分类;仅将干扰量提高到2%,分类错误的测试数据几乎达到总样本的一半;干扰量达到3%时,分类成功率已经不足40%,显然此时SVM分类器已经失效。

5.3.2 MNIST

在训练模型阶段,本文使用网格寻优设置多项式核-SVM模型的超参数。得到模型后,对测试集中的各个手写体数字的分类成功率单独测试,结果见表2。显然,所有手写体数字的总体分类成功率为95.77%;最低的类别为5,分类成功率也达到了92.71%,相对地,分类成功率最高的为手写体数字1,达到98.94%;超过总体分类成功率的包括数字0、1、4和6。

在不同扰动量下对各个类别的测试样本做扰动攻击,结果见表3。一方面,各类别对1%的少量扰动量不敏感,扰动效果最好的为数字3,它的分类成功率下降12.43%,最差的类别8仅仅下降0.73%。比较而言,3%~5%的扰动量会对大多数手写体数字造成显著影响,分类成功率平均下降接近50%。另一方面,手写体的各个数字对扰动量的敏感程度不同,如数字1在仅3%扰动量下就将分类成功率降至不足30%,5%的扰动量时攻击成功率达到99.83%,几乎完全分类错误;数字4在3%扰动量下的攻击成功率不如数字1,但也降至接近50%;相反地,一些数字在扰动量增至7%时仍表现一般,如数字2的分类成功率接近60%,这可能与数字2与其他数字的相似度高有关。数字2在扰动前的图像如图5(a)所示,在不同扰动量下的图像如图5(b)~图5(e)所示。

6 结束语

本文通过实验验证了提出的针对SVM的攻击样本生成方法。首先对人造随机数据集进行了

表2 测试集中各个手写体的分类准确率

总体准确率	0	1	2	3	4	5	6	7	8	9
95.77%	98.88%	98.94%	95.16%	95.74%	96.13%	92.71%	97.18%	94.65%	93.94%	93.76%

表3 不同扰动量下各类手写体的分类准确率

手写体数字	扰动前准确率	1%扰动量	3%扰动量	5%扰动量	7%扰动量
0	98.88%	97.49%	73.68%	36.41%	12.25%
1	98.94%	95.52%	29.38%	0.17%	0
2	95.16%	86.65%	75.40%	67.66%	56.87%
3	95.74%	83.31%	70.09%	52.14%	34.89%
4	96.13%	93.75%	50.97%	1.62%	0
5	92.71%	90.12%	61.52%	15.60%	4.55%
6	97.18%	95.68%	74.05%	34.07%	10.98%
7	94.65%	93.46%	77.54%	35.59%	18.82%
8	94.65%	93.92%	79.97%	39.60%	14.73%
9	93.94%	89.50%	49.56%	21.82%	8.18%

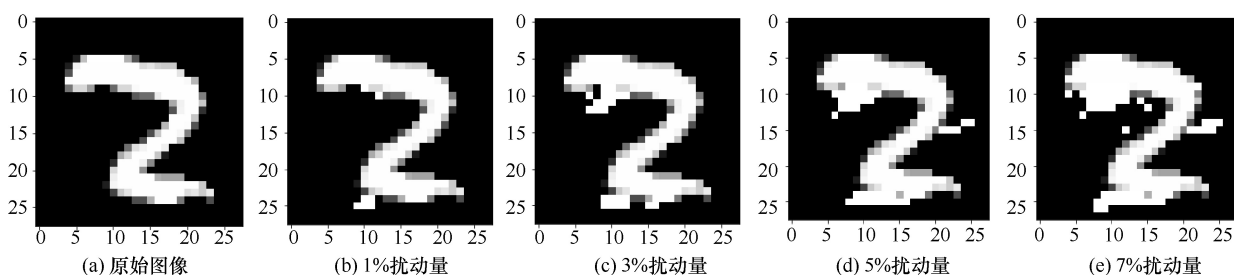


图5 不同扰动程度的图像示例

演示实验,证明了高维特征空间可以将输入空间中无法线性分类的数据进行较好的分类。在MNIST中,分别将不同的手写体数字单独作为一类,剩余的其他数字类别作为另一类进行实验,得到了令人满意的结果,如数字1在3%的扰动量下达到了超过70%的攻击成功率。本文提出的攻击样本生成方法从精确度上分析接近最优,但仍有提升的空间,如可以研究选择最优的扰动步长,这将是下一阶段的工作。

参考文献:

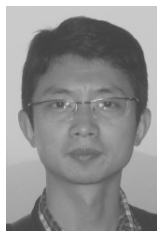
- [1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[Z]. 2014.
- [2] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//IEEE European Symposium on Security and Privacy, March 21, 2016, Saarbrücken, Germany. Piscataway: IEEE Press, 2016: 372-387.
- [3] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against deep learning systems using adversarial examples[J]. arXiv: 1602.02697, 2016.
- [4] ABDELWAHAB M, BUSSO C. Domain adversarial for acoustic emotion recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018: 1.
- [5] BARRENO M, NELSON B, SEARS R, et al. Can machine learning be secure?[C]//ACM Symposium on Information, Computer and Communications Security, March 21, 2006, Taipei, China. New York: ACM Press, 2006: 16-25.
- [6] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297.
- [7] 钱亚冠, 关晓惠, 吴淑慧, 等. 一种基于特征集构建的Bagging集成方法及其在流量分类中的应用[J]. 电信科学, 2018, 34(4): 41-48.
QIAN Y G, GUAN X H, WU S H, et al. An approach of Bagging ensemble based on feature set and application for traffic classification[J]. Telecommunications Science, 2018, 34(4): 41-48.
- [8] 李昊奇, 应娜, 郭春生, 等. 基于深度信念网络和线性单分类SVM的高维异常检测[J]. 电信科学, 2018, 34(1): 34-42.
- LI H Q, YING N, GUO C S, et al. High-dimensional outlier detection based on deep belief network and linear one-class SVM[J]. Telecommunications Science, 2018, 34(1): 34-42.
- [9] CLIFTON C W. Opportunities for private and secure machine learning[C]//ACM Workshop on Workshop on Aisec, Oct 27, 2008, Alexandria, USA. New York: ACM Press, 2008: 31-32.
- [10] RIAZI M S, WEINERT C, TKACHENKO O, et al. Chameleon: a hybrid secure computation framework for machine learning applications[J]. arXiv:1801.03239, 2018.
- [11] KIM M, SONG Y, WANG S, et al. Secure logistic regression based on homomorphic encryption: design and evaluation[J]. Jmir Medical Informatics, 2018, 6(2).
- [12] BIGGIO B, NELSON B, LASKOV P. Support vector machines under adversarial label noise[J]. Journal of Machine Learning Research, 2011, 20(3): 97-112.
- [13] BIGGIO B, NELSON B, LASKOV P. Poisoning attacks against support vector machines[C]//The 29th International Conference on Machine Learning, June 26-July 1, 2012, Edinburgh, Scotland. New York: ACM Press, 2012: 1467-1474.
- [14] LIN X, CHAN P P K. Causative attack to incremental support vector machine[C]//International Conference on Machine Learning and Cybernetics, July 13, 2014, Lanzhou, China. Piscataway: IEEE Press, 2014: 137-142.
- [15] LIU W G, HU Y G. Approach for attack signatures generating based on SVM and sequencealignment[J]. Journal of Central South University, 2012, 43(11): 4328-4332.
- [16] CHEN Z, TONDI B, LI X, et al. A gradient-based pixel-domain attack against SVM detection of global image manipulations[C]//Information Forensics and Security, Nov 10, 2018, Hong Kong, China. Piscataway: IEEE Press, 2018: 1-6.
- [17] AMRAEE S, VAFAEI A, JAMSHIDI K, et al. Abnormal event detection in crowded scenes using one-class SVM[J]. Signal Image & Video Processing, 2018(6): 1-9.
- [18] BENMAHAMED Y, TEGUAR M, BOUBAKEUR A. Application of SVM and KNN to Duval Pentagon 1 for transformer oil diagnosis[J]. IEEE Transactions on Dielectrics & Electrical Insulation, 2018, 24(6): 3443-3451.
- [19] SCHNALL A, HECKMANN M. Feature-space SVM adaptation for speaker adapted word prominence detection[J]. Computer Speech & Language, 2018.
- [20] ZHAO R, MAO K. Semi-random projection for dimensionality reduction and extreme learning machine in high-dimensional space[J]. IEEE Computational Intelligence Magazine, 2015,

10(3): 30-41.

[21] DEVORE R A, TEMLYAKOV V N. Some remarks on greedy algorithms[J]. Advances in Computational Mathematics, 1996, 5(1): 173-187.

[22] LECUN Y, CORTES C. The MNIST database of handwritten digits[Z]. 2010.

[作者简介]



钱亚冠（1976-），男，博士，浙江科技学院理学院副教授，主要研究方向为互联网流量分类、机器学习与大数据处理、对抗性机器学习。



吴淑慧（1975-），女，博士，浙江科技学院理学院讲师，主要研究方向为量子计算与机器学习。



云本胜（1980-），男，博士，浙江科技学院理学院讲师，主要研究方向为数据挖掘、服务计算。



关晓惠（1977-），女，浙江水利水电学院副教授，主要研究方向为对抗性机器学习。



任东晓（1982-），女，博士，浙江科技学院理学院高级工程师，主要研究方向为机器学习与大数据处理。