• 量子信息专栏 •

基于变分量子分类器的量子对抗攻击生成算法

侯晓凯[1]，吴热冰[2*]，王子竹[1*]，王晓霆[1*]

(1. 电子科技大学基础与前沿研究院 成都 610054；2. 清华大学自动化系 北京 海淀区 100084)

【摘要】量子分类器在扰动攻击下的脆弱性是量子机器学习中的基本理论问题之一。量子分类器的脆弱性是指其随着量子系统规模增大而更容易因为一些微小的扰动而分类错误的性质。这种微小扰动也被称为量子对抗攻击，而如何生成尽可能小的扰动使得量子分类器失效仍是一个开放问题。针对这一问题，提出了一种新的量子对抗攻击生成算法——量子混淆算法。该算法利用量子分类器关于输入数据的梯度信息来生成扰动，从而使得已训练好的量子分类器失效。数值仿真结果表明，与已有的量子对抗攻击方法相比，量子混淆算法可以通过更小的扰动实现对抗攻击，为理解分类器的有效性和脆弱性提供了新的思路。

关 键 词 量子对抗攻击； 量子分类器； 量子计算； 量子机器学习

中图分类号 TP301 文献标志码 A doi:10.12178/1001-0548.2023006

Quantum Adversarial Attack Generation Algorithm Based on Variational Quantum Classifiers

HOU Xiaokai[1], WU Rebing[2*], WANG Zizhu[1*], and WANG Xiaoting[1*]

(1. Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China Chengdu 610054;

2. Department of Automation, Tsinghua University Haidian Beijing 100084)

Abstract The vulnerability of quantum classifiers under adversarial attacks is one of the fundamental problems in quantum machine learning. The vulnerability of quantum classifiers refers to the property that a quantum classifier may be failed by small perturbations when the quantum system scales up. Such perturbations are also known as quantum adversarial attacks. How to generate small perturbations to fail a quantum classifier is still an open problem. To address this problem, we present a new quantum adversarial attack generation method, the quantum confounding algorithm, which generates perturbations that fail the trained quantum classifier by utilizing the gradient information of the quantum classifier with respect to the input data. Numerical results demonstrate that, compared with the existing quantum adversarial attack generation methods, our quantum confounding algorithm can generate significantly smaller perturbations that lead the quantum classifier to malfunction. This provides a new perspective in understanding the effectiveness and the vulnerability of quantum classifiers.

Key words quantum adversarial attack; quantum classifier; quantum computing; quantum machine learning

近年来，随着量子计算设备的不断发展与机器学习领域的不断探索，量子机器学习作为一个新兴的研究领域得到了广泛关注。一方面，依托于经典机器学习的强大能力，经典的机器学习算法已被应用于解决如量子纠缠判别[1-3]、量子纠错[4-6]与量子控制[7-9]等量子力学中的典型问题。另一方面，学者也着重于研究如何利用量子计算设备设计实现基于量子计算范式的量子机器学习算法，相比经典机器学习算法具有指数加速[10]。这些典型的量子机器学习算法包括量子主成分分析[11]、量子支持向量机[12]及量子玻尔兹曼机[13]等。

随着量子计算硬件进入中等规模含噪量子计算时代[14]，变分量子计算框架作为一种典型的经典–量子混合算法，利用经典计算机上运行的优化算法来对量子计算设备上线路的参数进行优化，从而可以解决特定的计算问题，如求解量子化学中的基态

收稿日期：2023－01－05；修回日期：2023－01－19

基金项目：国家重点研发计划 (2018YFA0306703，2021YFE0113100)；国家自然科学基金 (92265208，62173201，61833010)

作者简介：侯晓凯 (1995－)，男，主要从事量子计算与量子机器学习方面的研究.

*通信作者：吴热冰，E-mail：rbwu@tsinghua.edu.cn；王子竹，E-mail：zizhu@uestc.edu.cn；王晓霆，E-mail：xiaoting@uestc.edu.cn

能量[15-16]、求解组合优化问题[17-18] 等。特别地，学者提出了一系列基于变分量子计算框架的变分量子神经网络, 并从原理上验证了这种结构的量子神经网络与传统的神经网络在功能上是等价的[19]，甚至在一些特定问题上，量子神经网络展示出了超越传统神经网络的量子优势[20]。作为变分量子神经网络的一个重要应用，分类问题是机器学习领域，尤其是监督学习领域的一个重要分支。用于解决分类问题的变分量子神经网络也被称为量子分类器。这种典型的量子分类器包括量子卷积神经网络[21]、量子线路学习模型[22]、免复制量子神经网络[19] 及线路中心量子分类器[23]。这些量子分类器既可以用于解决量子多体系统基态所处的物理相的预测[19, 21]、量子态纠缠判别[24] 等典型的量子问题，也可以用于解决如鸢尾花分类[23]、手写数字图像识别[19, 23] 及信用卡评级分类[25] 等真实世界数据集的分类任务。

尽管量子分类器在近几年得到了广泛的关注和研究，但其中一个关键的基础理论问题，即如何解决量子分类器在扰动攻击下的脆弱性，依旧是一个量子机器学习领域的重要挑战。量子分类器的脆弱性是指，随着量子系统规模增大，量子分类器更容易因为一些微小的扰动而分类错误的性质。具体而言，随着问题规模的增大 (即实现量子分类器的量子比特数增加)，量子分类器的输入数据会不断靠近分类平面，从而一个微小的扰动就可以使得输入数据越过分类平面，进而使得量子分类器产生分类错误。已有研究表明，随着量子比特数量的增加，这种造成分类错误的扰动上界与量子系统对应的希尔伯特空间的维度成反比关系[26]。这种扰动也被称为量子对抗攻击。类比于经典机器学习中对抗样本生成的方法，针对量子分类器，现有文献已提出了一些量子对抗攻击生成方法，如量子自适应快速梯度符号法 (Q-FGSM) 和量子自适应基本迭代法 (Q-BIM)[27]。但如何能够更高效地生成量子对抗攻击，使得在生成扰动尽可能小的情况下，仍能令量子分类器发生错误，是一个亟待解决的问题。

为此，本文提出了一种新的量子对抗攻击生成算法——量子混淆算法，即通过一个迭代的过程来生成量子对抗攻击。在每一步迭代中，先计算量子分类器关于输入样本的梯度信息，并将这些信息处理为一个弱扰动，进而修改原始的输入数据。通过不断将这些弱扰动累加到原始数据中，最终可以得到关于原始数据的一个扰动，使得训练好的量子分类器发生分类错误。基于经典机器学习中典型的鸢尾花数据集进行了数值模拟，结果显示，对于一个对鸢尾花二分类能达到100%正确的量子分类器，通过利用量子混淆算法可以将分类准确率降为0%。将量子混淆算法与之前的 Q-FGSM 算法和 Q-BIM 算法进行了对比，结果显示，量子混淆算法可以生成更小的扰动使得量子分类器失效。

1　变分量子分类器

首先来简单回顾一下变分量子分类器及它的工作原理。变分量子分类器是一种典型的经典−量子混合算法，它主要包括一个量子计算机及一个经典优化器 (如图 1 所示)。对于一个输入数据$\boldsymbol{x}^{(i)}$，首先需要将其编码成一个量子态，然后利用量子计算机中的含参量子线路对其进行演化，并通过量子测量得到量子计算机关于样本的一个预测$\boldsymbol{u}^{(i)}$。根据一系列输入数据样本的预测值以及它们原始的标签$\boldsymbol{y}^{(i)}$，可以计算得到损失函数$\mathcal{L}(\boldsymbol{\theta})$和损失函数关于线路参数的梯度信息$\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$。进而，可以利用一个经典的优化器，通过迭代的方法不断更新线路参数$\boldsymbol{\theta}$以使得损失函数最小。使得损失函数最小的线路参数$\boldsymbol{\theta}^*$ 可以被认为是一组最优的线路参数，而对应的含参量子线路$\boldsymbol{U}(\boldsymbol{\theta}^*)$ 也可被认为是一个训练好的变分量子分类器。

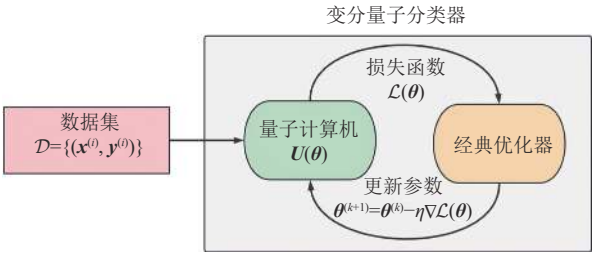

图 1　变分量子分类器示意图

具体来讲，对于一个给定的m类分类任务数据集$\mathcal{D}=\{(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})\}$，其中，样本数据$\boldsymbol{x}^{(i)}\in\mathbb{R}^N$ 是一个N维向量，其对应的标签$\boldsymbol{y}^{(i)}\in\mathbb{R}^m$是一个采用 one-hot 编码的$m$维向量，即如果$\boldsymbol{x}^{(i)}$属于第$k$类 ($k\in\{1,2,\cdots,m\}$)，则$\boldsymbol{y}^{(i)}$的第$k$个位置为1其余位置为0。那么，变分量子分类器中的量子计算机可以看成是一个函数映射：

$$u:\boldsymbol{x}\in\mathbb{R}^N\rightarrow u(\boldsymbol{x})\in\mathbb{R}^m \tag{1}$$

式中，$u(\boldsymbol{x})$ 满足$\|u(\boldsymbol{x})\|_1=1$。

为了实现这种变换，对于一个给定的原始输入数据样本$\boldsymbol{x}^{(i)}$，首先需要将它编码成一个对应的n 比特量子态$|\boldsymbol{x}^{(i)}\rangle$。通常, 所需的量子比特数满足

$N \leqslant 2^n$。本文使用一种量子机器学习中常见的信息编码方式，即振幅编码方法[28]，来将经典信息存储在量子系统中。具体来讲，对于一个N维的经典数据$\boldsymbol{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \cdots, x_N^{(i)}]^{\mathrm{T}}$，可以将其存储为如下形式的$n = \lceil \log_2 N \rceil$量子比特的量子态：

$$|\boldsymbol{x}^{(i)}\rangle = \frac{1}{\gamma}[x_1^{(i)}, x_2^{(i)}, \cdots, x_N^{(i)}, 0, \cdots, 0]^{\mathrm{T}} \tag{2}$$

式中，γ是一个归一化系数。在获得这样的量子态之后，通过量子计算机中的含参量子线路对其进行演化，这一过程可以被写成如下形式：

$$|\psi^{(i)}(\boldsymbol{\theta})\rangle = \boldsymbol{U}(\boldsymbol{\theta})|\boldsymbol{x}^{(i)}\rangle \tag{3}$$

式中，$\boldsymbol{U}(\boldsymbol{\theta})$是一个$2^n \times 2^n$的幺正矩阵，$\boldsymbol{\theta} \in \mathbb{R}^d$ 是一组含参量子线路的线路参数。在完成对量子态的演化之后，利用一组m个完备的 POVM 测量算子$\{\Pi_j\}_{j=1}^m$，对如式 (3) 所示的量子态进行量子测量。使用测量算子Π_j得到的测量结果即为数据$\boldsymbol{x}^{(i)}$被分为第j类的概率, 即：

$$u_j^{(i)} = \langle \Pi_j^{(i)} \rangle = \langle \psi^{(i)}(\boldsymbol{\theta}) | \Pi_j | \psi^{(i)}(\boldsymbol{\theta}) \rangle \tag{4}$$

在得到关于输入数据$\boldsymbol{x}^{(i)}$ 的分类预测值$\boldsymbol{u}^{(i)}$ 之后，可以结合它对应的样本标签$\boldsymbol{y}^{(i)}$ 来计算得到损失函数$\mathcal{L}(\boldsymbol{\theta})$。对于分类问题，一类常用的损失函数是交叉熵函数，对于二分类任务，交叉熵函数可以表示为：

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{M}\sum_i -[y^{(i)}\log(u^{(i)}) + (1-y^{(i)})\log(1-u^{(i)})] \tag{5}$$

式中，M表示样本的个数；对于多分类任务，交叉熵函数可以表示为：

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{1}{M}\sum_i \sum_{j=1}^m y_j^{(i)} \log\left(u_j^{(i)}\right) \tag{6}$$

更进一步，可以得到损失函数$\mathcal{L}(\boldsymbol{\theta})$ 关于变分量子线路中线路参数$\boldsymbol{\theta}$ 的梯度信息$\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})$，通过利用一个经典优化器，可以选择一种合适的优化算法对线路参数$\boldsymbol{\theta}$进行迭代训练，如随机梯度下降[29]、Adam[30]、BFGS[31] 及量子自然梯度下降[32] 等。经过充分多的迭代之后，可以得到最小化的损失函数以及对应的最优线路参数$\boldsymbol{\theta}^*$。此时，量子计算机$\boldsymbol{U}(\boldsymbol{\theta}^*)$ 可以看作一个训练好的变分量子分类器。

2 量子混淆算法

对于变分量子分类器，一个重要的基础理论问题是如何解决量子分类器在扰动攻击下的脆弱性。量子分类器的脆弱性是指，随着问题维度及系统规模的增大，量子数据在高维空间中更倾向于聚集在分类超平面的附近。已有研究表明，随着量子比特数的增加，量子数据与分类超平面的距离会指数减小[26]。这种距离的减小会造成一种现象，即微小的扰动就可以使得样本点越过分类超平面，从而令训练好的分类器产生分类错误。这种微小扰动的生成方法，也被称为量子对抗攻击生成算法。

受到经典对抗攻击生成算法的启发[33]，本文提出一种针对量子分类器的量子对抗攻击生成算法——量子混淆算法。量子混淆算法通过利用变分量子分类器关于编码原始数据量子态的梯度信息，来生成对抗攻击。具体来讲，一个训练好的变分量子分类器可被认为已经找到了关于分类数据的分类超平面。那么，可以通过迭代的方法，逐渐使得编码数据$\boldsymbol{x}^{(i)}$ 的量子态$|\boldsymbol{x}^{(i)}\rangle$ 向超平面移动，最终生成使得样本越过分类超平面的扰动。特别地，在第j次迭代中，根据受扰动后的量子态$|\tilde{\boldsymbol{x}}_j^{(i)}\rangle$在分类超平面上的正交投影得到一个弱扰动$r_j$，如下所示：

$$\boldsymbol{r}_j = -\frac{u(|\tilde{\boldsymbol{x}}_j^{(i)}\rangle)}{\|\nabla_{\boldsymbol{x}} u(|\tilde{\boldsymbol{x}}_j^{(i)}\rangle)\|_2^2} \nabla_{\boldsymbol{x}} u(|\tilde{\boldsymbol{x}}_j^{(i)}\rangle) \tag{7}$$

式中，u表示一个训练好的量子变分分类器；$\nabla_{\boldsymbol{x}} u(|\tilde{\boldsymbol{x}}_j^{(i)}\rangle)$ 表示量子分类器的输出关于$|\tilde{\boldsymbol{x}}_j^{(i)}\rangle$的梯度。更进一步，引入一个扰动系数$\epsilon$ 来约束每次生成的扰动规模。这样，可以得到对于样本$\boldsymbol{x}^{(i)}$ 生成对抗攻击$\boldsymbol{r}^{(i)}$ 在第j 步的累积更新规则：

$$\boldsymbol{r}^{(i)} \leftarrow \boldsymbol{r}^{(i)} + \epsilon \boldsymbol{r}_j \tag{8}$$

在获得第j步迭代中生成的扰动$\boldsymbol{r}^{(i)}$ 以后，将扰动施加在样本$|\tilde{\boldsymbol{x}}_j^{(i)}\rangle$ 上并检查施加扰动后的数据样本能否使得变分量子分类器发生分类错误。如果可以，则输出这样的扰动，如果不可以就进入下一次的迭代更新。将整个扰动的生成过程总结在算法 1 中。

算法 1　量子混淆算法

输入：量子分类器u，数据样本$(|\boldsymbol{x}^{(i)}\rangle, \boldsymbol{y}^{(j)})$，扰动系数$\epsilon$，最大迭代次数 N_{Iter}

输出：样本的对抗扰动 $\boldsymbol{r}^{(i)}$

$j \leftarrow 0$, $\boldsymbol{r}_j \leftarrow 0$, $\boldsymbol{r}^{(i)} \leftarrow 0$, $|\tilde{\boldsymbol{x}}_j^{(i)}\rangle \leftarrow |\boldsymbol{x}^{(i)}\rangle$

while: $\arg\max(u(|\tilde{\boldsymbol{x}}_j^{(i)}\rangle)) == \arg\max(\boldsymbol{y}^{(i)})$

and $j \leqslant N_{\mathrm{Iter}}$ do

$$\boldsymbol{r}_j \leftarrow -\frac{u(|\tilde{\boldsymbol{x}}_j^{(i)}\rangle)}{\|\nabla_{\boldsymbol{x}} u(|\tilde{\boldsymbol{x}}_j^{(i)}\rangle)\|_2^2} \nabla_{\boldsymbol{x}} u(|\tilde{\boldsymbol{x}}_j^{(i)}\rangle)$$

$$\boldsymbol{r}^{(i)} \leftarrow \boldsymbol{r}^{(i)} + \epsilon r_j$$

$$|\tilde{\boldsymbol{x}}_j^{(i)}\rangle \leftarrow \frac{1}{\|\tilde{\boldsymbol{x}}_j^{(i)}\rangle + \boldsymbol{r}^{(i)}\|_2}(|\tilde{\boldsymbol{x}}_j^{(i)}\rangle + \boldsymbol{r}^{(i)})$$

$$j \leftarrow j+1$$

end

Return $\boldsymbol{r}^{(i)}$

3　应　用

通过数值仿真验证了量子混淆算法的有效性，进一步将量子混淆算法与之前提出的 Q-FGSM 算法和 Q-BIM 算法进行了比较。数值仿真是基于鸢尾花数据集的二分类任务进行的[25]。具体来讲，使用了两种鸢尾花 (Setosa、Versicolour) 的数据作为实验样本。这些实验样本总共有 100 条数据，每一条数据包括了萼片长度、萼片宽度、花瓣长度和花瓣宽度 4 个维度的属性。对于每条样本对应的标签，使用 one-hot 编码，即如果样本属于 Setosa 类别，则对应的标签是$[1,0]^{\mathrm{T}}$, 如果样本属于 Vesicolour 类别，其对应的标签是$[0,1]^{\mathrm{T}}$。

为了训练量子分类器，使用了振幅编码的方案将这些经典数据转换成对应的量子态。对于样本$\boldsymbol{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, x_4^{(i)}]$，将其编码到一个两比特量子态$|\boldsymbol{x}^{(i)}\rangle$，如下所示：

$$|\boldsymbol{x}^{(i)}\rangle = \frac{1}{\|\boldsymbol{x}^{(i)}\|_2}\sum_{j=1}^{4} x_j^{(i)}|j\rangle \tag{9}$$

式中，$\{|j\rangle\}_{j=1}^4$ 表示由两个量子比特组成的希尔伯特空间中一组完备的计算基底。

为了得到一个能够分辨出样本数据类别的变分量子分类器，采用了一种如图 2 所示的变分量子分类器[23]。这种量子分类器由 3 个量子比特组成。前两个比特用来存储输入的量子态，第三个比特用来存储量子分类器的预测标签。整个变分量子分类器由L 层变分量子线路组成。每层变分量子线路包含了若干含参单比特量子门和两比特量子门。在这里，使用的含参单比特量子门包括绕 x 轴旋转的单比特门$R_x(\theta) = \mathrm{e}^{-i\frac{\theta}{2}\sigma_x}$ 以及绕 z 轴旋转的单比特量子门$R_z(\theta) = \mathrm{e}^{-i\frac{\theta}{2}\sigma_z}$。其中，$\sigma_x$和$\sigma_z$分别表示泡利 X 矩阵与泡利 Z 矩阵。至于两比特量子门，采用的是 CNOT 门，其数学形式可以写成：

$$\mathrm{CNOT} = |0\rangle\langle 0|\otimes\mathbb{I} + |1\rangle\langle 1|\otimes\sigma_x \tag{10}$$

式中，$\mathbb{I}$表示一个2×2的单位矩阵。特别地，在实验中，使用了一个$L=5$的变分量子线路。

本文采用的变分量子分类器是一个基于三比特 L 层的变分量子线路。其中每一层包含了 9 个单比特旋转门和两个 CNOT 门。在执行完所有的量子门之后，对第三个比特进行量子测量，测量其关于泡利 Z 矩阵的期望值。

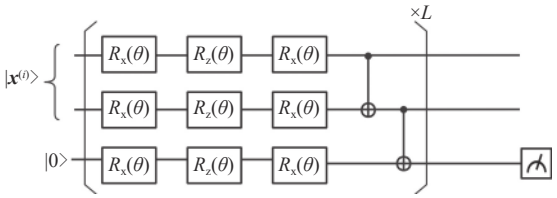

图 2　变分量子分类器线路结构示意图

对于这样的变分量子分类器，在线路末尾测量第三个比特关于泡利 Z 矩阵的期望值$\langle\sigma_z\rangle$，并将$u = \dfrac{\langle\sigma_z\rangle + 1}{2}$ 作为量子分类器的输出值。由于$0 \leqslant u \leqslant 1$，可以将$u$ 作为样本属于 Setosa 类的概率，将$1-u$ 作为 Vesicolour 类别的概率。因此，对于一个输入样本$\boldsymbol{x}^{(i)}$，量子分类器的输出即为$[u, 1-u]^{\mathrm{T}}$。更进一步，将二分类交叉熵函数 (见式 (5)) 作为优化的目标函数$\mathcal{L}(\boldsymbol{\theta})$, 并通过 Adam 算法对线路参数进行优化。通过训练，实现的变分量子分类器对数据的分类准确率可以达到100%。

进一步利用量子混淆算法，Q-FGSM 算法和 Q-BIM 算法生成关于数据集的对抗攻击，使得分类器产生分类错误。为了衡量 3 种量子对抗攻击算法的表现，将原始数据$|\boldsymbol{x}^{(i)}\rangle$以及其添加扰动后对应的量子态$|\tilde{\boldsymbol{x}}^{(i)}\rangle$之间的保真度作为对所生成扰动大小的度量，即保真度越高表示生成的扰动越小。对于样本$|\boldsymbol{x}^{(i)}\rangle$的保真度，其定义为：

$$f(|\boldsymbol{x}^{(i)}\rangle, |\tilde{\boldsymbol{x}}^{(i)}\rangle) = |\langle\boldsymbol{x}^{(i)}|\tilde{\boldsymbol{x}}^{(i)}\rangle|^2 \tag{11}$$

更进一步，可以定义在整个数据集$\mathcal{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^M$上的平均保真度$F$为：

$$F = \frac{1}{M}\sum_{i=1}^{M} f(|\boldsymbol{x}^{(i)}\rangle, |\tilde{\boldsymbol{x}}^{(i)}\rangle) \tag{12}$$

除此之外，在使得量子分类器失效 (即分类正确率为0%) 时，定义了平均扰动比例R来衡量每种量子对抗攻击生成算法所生成的扰动规模与原始数据大小的比例，其定义如下：

$$R = \frac{1}{M}\sum_i \frac{\|\boldsymbol{r}^{(i)}\|_2}{\|\boldsymbol{x}^{(i)}\|_2} \tag{13}$$

式中，M 表示数据集的样本数量。仿真结果如图 3 及表 1 所示，可以看到，在使得训练好的量子分类

器失效的情况下，量子混淆算法生成的平均扰动比例为0.192 2， 远小于 Q-FGSM 的平均扰动比例1.380 0 与 Q-BIM 的平均扰动比例0.311 0。与此同时，利用量子混淆算法生成的使得分类器完全失效的对抗样本$\{|\tilde{\boldsymbol{x}}^{(i)}\rangle\}$与原始的数据样本$\{|\boldsymbol{x}^{(i)}\rangle\}$之间的平均保真度为0.962 6，高于 Q-FGSM 的平均保真度0.487 5 与 Q-BIM 的平均保真度0.918 2。实验结果表明，相比于 Q-FGSM 与 Q-BIM 算法，利用量子混淆算法可以生成更小的扰动使得量子分类器完全失效。由此可见，量子混淆算法是一种高效的量子对抗攻击生成算法。

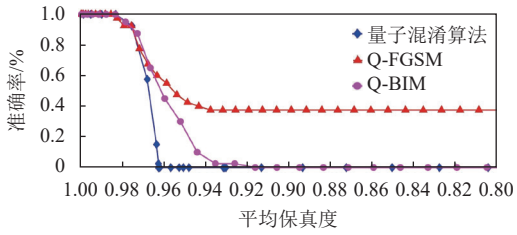

图 3 变分量子分类器基于样本点的分类准确率与平均保真度的关系

表 1 量子对抗攻击生成算法在分类器失效时的平均扰动比例与对抗样本平均保证度

算法	平均扰动比例R	平均保真度F
量子混淆算法	0.192 2	0.962 6
Q-FGSM	1.380 0	0.487 5
Q-BIM	0.311 0	0.918 2

4 结 束 语

本文提出了一种针对变分量子分类器的量子对抗攻击生成方法——量子混淆算法。该算法通过迭代的方法生成一种微小扰动。将这种扰动施加到编码了原始数据的量子态上可以生成原始数据对应的对抗样本，这些对抗样本在保持与原始样本较高保真度的情况下，可以使得训练好的分类器完全失效。通过仿真实验，证明了量子混淆算法可以通过利用量子分类器关于输入样本点的梯度信息来高效生成样本的扰动，与之前提出的两种量子对抗攻击生成方法 (Q-FGSM 算法和 Q-BIM 算法) 相比，本文提出的量子混淆算法在同等效果的情况下，可以生成更小的扰动，进而生成与原始数据相比保真度更高的对抗样本。综上所述，量子混淆算法是一种高效的量子对抗攻击生成方法，为理解量子分类器的有效性和脆弱性提供了新思路。

参 考 文 献

[1] LEVINE Y, YAKIRA D, COHEN N, et al. Deep learning and quantum entanglement: Fundamental connections with implications to network design [EB/OL]. (2017-04-10). http://doi.org/10.48550/arXiv.1704.01552.

[2] LU S R, HUANG S L, LI K R, et al. Separability-Entanglement classifier via machine learning[J]. Physical Review A, 2018, 98(1): 012315.

[3] GAO J, QIAO L F, JIAO Z Q, et al. Experimental machine learning of quantum states[J]. Physical Review Letters, 2018, 120(24): 240501.

[4] FÖSEL T, TIGHINEANU P, WEISS T, et al. Reinforcement learning with neural networks for quantum feedback[J]. Physical Review X, 2018, 8(3): 031084.

[5] BAIREUTHER P, O'BRIEN T E, TARASINSKI B, et al. Machine-Learning-Assisted correction of correlated qubit errors in a topological code[J]. Quantum, 2018, 2: 48.

[6] NAUTRUP H P, DELFOSSE N, DUNJKO V, et al. Optimizing quantum error correction codes with reinforcement learning[J]. Quantum, 2019, 3: 215.

[7] BUKOV M, DAY A G, SELS D, et al. Reinforcement learning in different phases of quantum control[J]. Physical Review X, 2018, 8(3): 031086.

[8] NIU M Y, BOIXO S, SMELYANSKIY V N, et al. Universal quantum control through deep reinforcement learning[J]. NPJ Quantum Information, 2019, 5(1): 1-8.

[9] ZENG Y X, SHEN J, HOU S C, et al. Quantum control based on machine learning in an open quantum system[J]. Physics Letters A, 2020, 384(35): 126886.

[10] BIAMONTE J, WITTEK P, PANCOTTI N, et al. Quantum machine learning[J]. Nature, 2017, 549(7671): 195-202.

[11] LLOYD S, MOHSENI M, REBENTROST P. Quantum algorithms for supervised and unsupervised machine learning. [EB/OL]. (2013-11-04). http://doi.org/10.48550/arXiv.1307.0411.

[12] REBENTROST P, MOHSENI M, LLOYD S. Quantum support vector machine for big data classification[J]. Physical Review Letters, 2014, 113(13): 130503.

[13] AMIN M H, ANDRIYASH E, ROLFE J, et al. Quantum boltzmann machine[J]. Physical Review X, 2018, 8(2): 021050.

[14] PRESKILL J. Quantum computing in the NISQ era and beyond[EB/OL]. (2018-07-30). http://doi.org/10.48550/arXiv.1801.00862.

[15] KANDALA A, MEZZACAPO A, TEMME K, et al. Hardware-Efficient variational quantum eigensolver for small molecules and quantum magnets[J]. Nature, 2017, 549(7671): 242-246.

[16] WANG D, HIGGOTT O, BRIERLEY S. Accelerated variational quantum eigensolver[J]. Physical Review Letters, 2019, 122(14): 14050.

[17] FARHI E, GOLDSTONE J, GUTMANN S. A quantum approximate optimization algorithm[EB/OL]. (2014-11-14). http://doi.org/10.48550/arXiv.1411.4028.

[18] GUERRESCHI G G, MATSUURA A Y. QAOA for max-cut requires hundreds of qubits for quantum speed-up[J].

Scientific Reports, 2014, 9(1): 1-7.
[19] HOU X, ZHOU G, LI Q, et al. A duplication-free quantum neural network for universal approximation[EB/OL]. (2022-11-21). http://doi.org/10.48550/arXiv.2211.11228.
[20] HUANG H Y, BROUGHTON M, MOHSENI M, et al. Power of data in quantum machine learning[J]. Nature Communications, 2022, 12(2631): 1-9.
[21] CONG I, CHOI S, LUKIN M D. Quantum convolutional neural networks[J]. Nature Physics, 2019, 15(12): 1273.
[22] MITARAI K, NEGORO M, KITAGAWA M, et al. Quantum circuit learning[J]. Physical Review A, 2018, 98(3): 032309.
[23] SCHULD M, BOCHAROV A, SVORE K M, et al. Circuit-Centric quantum classifiers[J]. Physical Review A, 2020, 101(3): 032308.
[24] MA Y C, YUNG M H. Transforming Bell's inequalities into state classifiers with machine learning[J]. NPJ Quantum Information, 2018, 4: doi: 10.1038/s41534-018-0081-3.
[25] KILLORAN N, BROMLEY T R, ARRAZOLA J M, et al. Continuous-Variable quantum neural networks[J]. Physical Review Research, 2019, 1(3): 033063.
[26] LIU N, WITTEK P. Vulnerability of quantum classification to adversarial perturbations[J]. Physical Review A, 2020, 101(6): 062331.
[27] LU S, DUAN L M, DENG D L. Quantum adversarial machine learning[J]. Physical Review Research, 2020, 2(3): 033212.
[28] NAKAJI K, UNO S, SUZUKI Y, et al. Approximate amplitude encoding in shallow parameterized quantum circuits and its application to financial market indicators[J]. Physical Review Research, 2022, 4(2): 023136.
[29] BOTTOU L. Stochastic gradient descent tricks[M]// MONTAVON G, ORR G B, MÜLLER K R. Neural Networks: Tricks of the Trade, Heidelberg: Springer, 2012.
[30] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2014-12-22). http://doi.org/10.48550/arXiv.1412.6980.
[31] LIU D C, NOCEDAL J. On the limited memory BFGS method for large scale optimization[J]. Mathematical Programming, 1989, 45(1): 503-528.
[32] STOKES J, IZAAC J, KILLORAN N, et al. Quantum natural gradient[J]. Quantum, 2020, 4: 269.
[33] MOOSAVI D S, FAWZI A, FROSSARD P. DeepFool: A simple and accurate method to fool deep neural networks[C]//In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 1-9.

编　辑　蒋　晓