arXiv:2305.02396v2 [cs.LG] 9 Aug 2023

特征工程能否助力量子机器学习进行恶意软件检测?

Ran Liu马里兰大学巴尔的摩县分校 {rliu2@umbc.edu}

马克西姆·埃伦
马里兰大学巴尔的摩县分校
meren1@umbc.edu

Charles Nicholas马里兰大学巴尔的摩县分校 nicholas@umbc.edu

摘要

随着恶意软件攻击数量不断增加、手段日益复杂，基于机器学习（ML）的恶意软件检测系统变得愈发重要。与此同时，许多用于恶意软件分类的热门机器学习模型都是监督式方法。这些监督式分类器往往难以很好地泛化到新型恶意软件。因此，为了检测新的恶意软件样本，必须频繁重新训练这些模型，这可能耗时较长。我们的工作通过一种理论量子机器学习的混合框架来解决这一问题，并结合特征选择策略，以缩减数据规模和恶意软件分类器的训练时间。

1 引言

随着恶意软件攻击数量不断增加、手段日益复杂，基于机器学习（ML）的恶意软件检测系统变得愈发重要。与此同时，许多用于恶意软件分类的热门机器学习模型都是监督式方法。这些监督式分类器往往难以很好地泛化到新型恶意软件。因此，为了检测新的恶意软件样本，必须频繁重新训练这些模型，这可能耗时较长。我们的工作通过一种理论量子机器学习的混合框架来解决这一问题，并结合特征选择策略，以缩减数据规模和恶意软件分类器的训练时间。

从理论上讲，量子并行性可以为许多机器学习算法带来潜在的加速。已有研究表明，一个 N 维数据集可以映射到使用 logN 个量子比特表示的量子态上。因此，经典数据的映射过程需要 O(logN) 步骤[1]。此外，量子操作在某些特定步骤上也表现出优势，其中量子傅里叶变换和矩阵求逆的时间复杂度为 O(poly(log N))，距离估计和内积的时间复杂度为 O(log N)[2]。已有研究还表明，在聚类问题中使用量子算法具有优势。Harrow 等人利用量子算法进行聚类，实现了 O(log(MN)) 的运行时间复杂度；相比之下，最佳的经典聚类算法需要 O(poly(MN))[3]。更具体地说，当前关于量子退火的研究表明，其在某些特定的机器学习算法上具有优势，并具备部分抗噪能力；当数据集较大时，这限制了其在一些实际问题中的应用。在本研究中，我们探索了一种基于经典机器学习方法的特征选择策略，以便在量子门模型机器（例如 IBM 的量子处理器）上训练分类器，适用于量子比特容量非常有限的机器。我们将使用一些最先进的量子算法对该方法进行检验，并在量子门模型机器上运行这些算法。

2 个量子比特

“Qubit”或“Quantum bit”是量子计算中的基本对应概念，定义于希尔伯特空间中。使用狄拉克记号，给定希尔伯特空间中的一组正交归一基，量子比特可以按一定概率测量为某个基态。我们可以使用状态向量语言来表示量子系统。例如，我们可以生成叠加态

通过对计算基 $|0\rangle$ 和 $|1\rangle$中的状态进行线性组合:

$$|\varphi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{1}$$

其中 $\alpha, \beta \in \mathbb{C}$ 和 $|\alpha|^2 + |\beta|^2 = 1$。不同于只能处于状态 0 或 1 的经典比特，量子比特在被观测之前可以处于连续的基态中。例如，1 表示量子比特可以以 $|\alpha|^2$ 的概率处于状态 $|0\rangle$，或以 $|\beta|^2$ 的概率处于状态 $|1\rangle$。

3 变分量子分类器

在我们的研究中，我们使用 Havlíˇcek, V., Córcoles, A.D., Temme, K. 等人提出的变分量子分类器 [4]。区间 $(0, 2\pi]$ 内的一个经典数据点，首先通过如下构造的非线性电路映射到布洛赫球上:

$$\mathcal{U}_\Phi(\vec{x}) = U_{\Phi(\vec{x})} H^{\otimes n} U_{\Phi(\vec{x})} H^{\otimes n}$$

H 是哈达玛门，而

$$U_{\Phi(\vec{x})} = \exp\left(i \sum_{S \subseteq [n]} \phi_S(\vec{x}) \prod_{i \in S} Z_i \right)$$

然后构建用于优化的变分电路，称为*l*层电路。*l*表示将构建的变分电路重复*l* 次。接下来，对输出比特串应用 Z基测量 f: $\{0,1\}^n \longrightarrow \{+1, -1\}$。最后，重复测量 n 次以得到概率分布，并为输入向量分配标签。

4 相关工作

支持向量机（SVM）算法是一种监督学习算法，用于解决二分类问题。SVM 通过寻找一个超平面来区分属于两个不同类别的样本。超平面定义如下:

$$\vec{w} \cdot \vec{x} + b = 0$$

其中 $\vec{x}$ 是法向量，而 b 是偏移量。因此，对于类别 A 和 B，训练良好的 SVM 超平面可以给出:

$$\vec{w} \cdot \vec{x} + b \geq 1 if \vec{x} \in A$$

$$\vec{w} \cdot \vec{x} + b \leq -1 if \vec{x} \in B$$

因此，SVM 的问题是在满足以下条件的情况下，最大化类别 A、B 之间的间隔:

$$y\left(\vec{w} \cdot \vec{x} + b\right) \geq 1$$

新向量 $\vec{x_0}$ 的分类结果由下式给出:

$$sign\left(\vec{w} \cdot \vec{x_0} + b\right)$$

Havlíˇcek, V., Córcoles, A.D., Temme, K. 等人 [4]提出了两种支持向量机类型的分类器，将希尔伯特空间用作特征空间，以实现量子优势。第一种分类器旨在构建 [5] [6] 中讨论的变分电路，从而在希尔伯特空间中生成超平面。另一种分类器则利用量子计算机直接估计希尔伯特空间的核函数。如果传统计算机难以模拟这些电路，量子机制便能带来量子优势。他们的特征映射策略可通过图 1 直观呈现。区间 $(0, 2\pi]$ 内的经典数据点首先通过以下非线性电路映射到布洛赫球上:

$$\mathcal{U}_\Phi(\vec{x}) = U_{\Phi(\vec{x})} H^{\otimes n} U_{\Phi(\vec{x})} H^{\otimes n}$$

H 是哈达玛门，并且

$$U_{\Phi(\vec{x})} = \exp\left(i \sum_{S \subseteq [n]} \phi_S(\vec{x}) \prod_{i \in S} Z_i \right)$$

在他们的第一种方法中，首先使用上述特征映射策略将经典数据映射到希尔伯特空间。接着，他们构建变分电路以进行优化，并将其称为*l*层电路。*l* 表示重复构建的变分电路*l* 次。随后，对输出比特串应用Z基测量 f: $\{0,1\}^n \longrightarrow \{+1, -1\}$。最后，重复测量 n 次，以得到概率分布。

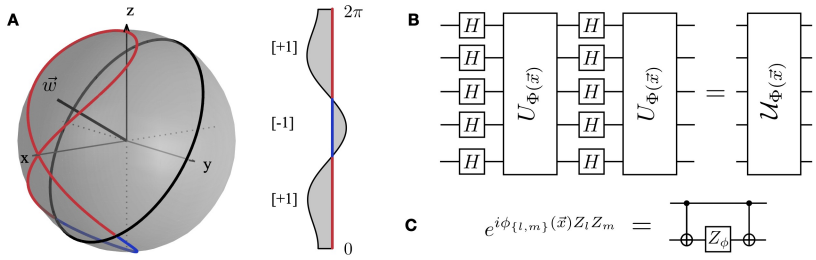


图 1：区间 $(0, 2\pi]$ in A 中的一个经典数据点，通过 B 中的非线性电路映射到布洛赫球上 [7]

并为输入向量分配标签。在他们的第二种方法中，量子计算机被直接用于估计。首先，量子计算机参与训练过程，以估计训练数据集中所有数据对的核。然后，在整理好支持向量机的优化问题后，再次使用量子计算机估计一个新数据点与所有支持向量之间的核。他们使用生成的定制数据测试了这两种方法。结果表明，即使存在噪声，成功率仍达到100%。

Patrick Rebentrost, Ma soud Mohseni 等人 [8] 提出了一种量子方法，用于分析支持向量机的最小二乘近似 n 支持向量机。在高层面上，他们尝试通过求解线性方程来获取优化参数：

$$F\left(b\right)\vec{a} \equiv \left(0\right)\&1I\&K + \gamma^{-1}I_N\left(b\right)\vec{a} = \left(0\right)\vec{y}$$

其中，K 是训练数据集中成对样本的核，a 是权重，b 是基向量，$\vec{y} = (y_1、\ldots\ldots、y_M)^T$ 是标签。因此，矩阵 F 为 $(M+1)\times(M+1)$ 他们的方法可概括如下：首先，将经典数据点映射到希尔伯特空间：

$$|\vec{x_i}\rangle = \frac{1}{|\vec{x_i}|}\sum_{i=1}^{M}(\vec{x_i})_j|j\rangle$$

其次，应用 HHL算法 [9] 以获得优化超平面参数 $\vec{a}$ 和 b。然后，将量子寄存器转换为：

$$|b,\vec{a}\rangle = \frac{1}{\sqrt{C}}(b|0\rangle + \sum_{k=1}^{M}(a_k|k\rangle)$$

其中 $C = b^2 + \sum_{k=1}^{M} a_k^2$。分类结果为：

$$y(\vec{x_0}) = (\sum_{i=1}^{M} a_i(\vec{x_i}\cdot\vec{x_0}) + b)$$

5 实验

在我们的实验中，我们使用 Drebin数据集，其中包含 +15k 个良性和恶意软件样本，并具有 215 个基于动态分析和静态分析的特征[10]。我们使用 XGBoost 和决策树（DT）识别出最重要的 20 个特征。随后，我们利用这些特征，在 IBM Qiskit模拟器上使用变分量子分类器（VQC）开展实验。初步结果表明，使用 XGBoost 选出的特征时，VQC 在 10,000 个样本、测试集占比为 50% 的情况下，测试准确率达到 78.91%。相比之下，使用 DT 进行特征选择的 VQC 测试准确率为 62.41%。我们还使用 IBM 的 5 量子比特机器进行了测试。本轮实验受到 IBM量子计算机量子比特数量的限制。因此，我们进行了十次实验，每次随机选取 20 个样本，测试集占比均为 50%。为表明结果具有统计显著性，我们使用量子支持向量机（QSVM），以 95% 置信区间（CI）报告最终准确率。使用 XGBoost 选出的特征训练的模型，最终平均准确率为 74%（+- 11.35%）。

接下来，我们探索了基于经典机器学习方法的过采样和欠采样策略，以帮助在量子门模型机器上训练用于恶意软件检测的分类器。我们的初步实验结果表明：

- 采用过采样技术时，使用 SMOTE 后，准确率从 74% 提高至 80.15%；使用自适应合成（ADASYN）后，准确率提高至 78.06%。

- 采用欠采样技术时，通过 K-均值移除样本，准确率从 74% 提高至 78.62%；移除与其邻域“不够”一致的样本后，准确率提高至 77.26%。
- 为克服 SMOTE 引入的噪声样本，我们结合使用过采样和欠采样，使准确率提升至 83.78%。
- 我们在 IBM 的 5 量子比特量子计算机上进行了实验。由于量子计算机可用的量子比特数量有限，我们使用了 20 个样本，并按 50% 的比例划分测试集。量子支持向量机的准确率为 56%，变分量子分类器在恶意软件检测中的成功率达到 80%。

6 结论

本研究探讨了特征工程和量子计算机在恶意软件检测任务中的应用。受物理限制影响，量子计算机在恶意软件检测中的实际应用仍处于早期阶段。我们希望本研究能激励其他研究人员探索这一方向。

参考文献

[1] Vittorio Giovannetti、Seth Lloyd 和 Lorenzo Maccone。量子随机存取存储器。2007年8月。[2] Michael A. Nielsen 和 Isaac L. Chuang。《量子计算与量子信息》。剑桥大学出版社，2012年6月。[3] Aram W. Harrow、Avinatan Hassidim 和 Seth Lloyd。用于求解线性方程组的量子算法。2008年11月。[4] Vojtˇech Havlíˇcek、Antonio D. Córcoles、Kristan Temme、Aram W. Harrow、Abhinav Kandala、Jerry M. Chow 和 Jay M. Gambetta。利用量子增强特征空间进行监督学习。《自然》，567(7747):209–212，2019年3月。[5] Kosuke Mitarai、Makoto Negoro、Masahiro Kitagawa、Keisuke Fujii。量子电路学习，2018年。[6] Edward Farhi、Hartmut Neven。在近期处理器上使用量子神经网络进行分类，2018年。[7] Vojtech Havlicek、Antonio D. Córcoles、Kristan Temme、Aram W. Harrow、Abhinav Kandala、Jerry M. Chow、Jay M. Gambetta。利用量子增强特征空间进行监督学习，2018年。[8] Patrick Rebentrost、Masoud Mohseni、Seth Lloyd。用于大数据分类的量子支持向量机，2014年。[9] Aram W. Harrow、Avinatan Hassidim、Seth Lloyd。用于求解线性方程组的量子算法，2009年。[10] Suleiman Y. Yerima 和 Sakir Sezer。DroidFusion：一种用于安卓恶意软件检测的新型多层分类器融合方法。《IEEE控制论汇刊》，49(2):453–466，2019年2月。