DOI: 10.3969/j.issn.1007-5461.2024.01.011

基于参数化角编码的量子 *K*-means 算法

冯微军， 郭躬德， 林 崧*

(福建师范大学计算机与网络空间安全学院， 福建 福州 350007)

摘 要: 结合 *K*-means 算法和角编码技术, 提出了一种无需量子随机存储 (QRAM) 的量子 *K*-means 算法。该算法利用量子操作的并行性, 仅需对数数量的时间复杂度就能完成数据的加载; 并且通过对输入数据进行参数预处理操作, 确定数据分量的参数阈值, 解决了样本不同特征尺度差异的问题。该算法由编码数据、相似度度量、量子最小值搜索和质心迭代更新四个主要步骤组成, 细致描述了这些步骤所涉及的算子和线路构建, 并对关键线路进行了仿真模拟。实验结果和经典预测结果一致, 验证了所提量子 *K*-means 算法的可靠性。此外, 理论分析表明所提出算法相比于经典算法在运行时间上有平方级加速。

关 键 词: 量子光学; 量子 *K*-means 算法; 角编码; 量子相位估计; 多量子比特交换测试

中 图 分 类 号 : TP319 **文 献 标 识 码 :** A **文章编号:** 1007-5461(2024)01-00113-12

Quantum *K*−means algorithm based on parameterized angle encoding

*FENG Weijun, GUO Gongde, LIN Song**

(College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350007, China)

Abstract: A quantum *K*-means algorithm without quantum random access memory (QRAM) is proposed by combining *K*-means algorithm and angle encoding technology. This algorithm makes use of parallel quantum operations and can complete data loading with only logarithmic time complexity. And by pre-processing the input data, the parameter threshold of the data components is determined, so the problem of different characteristic scales of samples can be solved according to the algorithm. The main body of the algorithm consists of four main steps: coding data, similarity measurement, quantum minimum search and centroid iterative update. The operators and circuit construction involved in these steps are described in detail. Numerical experiments based on the proposed circuit show that the results of the proposed algorithm are consistent with the classical prediction results, verifying the reliability of the quantum *K*-means algorithm combined with parameters. In addition, theoretical analysis shows that the proposed algorithm has square acceleration in running time compared with the classical algorithms.

Key words: quantum optics; quantum *K*-means algorithm; angle encoding; quantum phase estimation; multi-qubits swap-test

基金项目: 国家自然科学基金 (62171131, 61976053, 61772134), 福建省高等学校新世纪优秀人才支持计划, 福建省自然科学基金 (2018J01776)

作者简介: 冯微军 (1996 -), 浙江温州人, 研究生, 主要从事量子机器学习方面的研究。E-mail: 1518760342@qq.com

导师简介: 郭躬德 (1965 -), 福建龙岩人, 博士, 教授, 博士生导师, 主要从事数据挖掘、量子机器学习等方面的研究。 E-mail: ggd@fjnu.edu.cn

收稿日期: 2022-03-29; **修改日期:** 2022-06-20

*通信作者。 E-mail: lins95@fjnu.edu.cn

0 引 言

现代信息产业的高速发展以及数据的爆炸增长, 让人们对计算力的需求远远超过以往任何一个时代。IDC DataAge 2025 白皮书显示, 全球数据量总和预计到 2025 年将达到 175 ZB, 因此, 数据分析迎来了巨大的挑战。早在 1982 年, Feynman[1]提出了量子模拟的构想, 开创了量子计算这种本质上全新的计算模型。随后, Lloyd 等[2] 提出了第一个量子哈密顿模拟算法, 证实了 Feynman 的构想。1985 年, Deutsch[3]提出通用容错量子计算机, 描述了图灵机的量子泛化, 证明了量子理论和通用计算机的理论是相容的, 并且可能比传统计算机具有更强的计算能力。在探寻量子优势的过程中, Deutsch[4]在 1989 年首次提出了 Deutsch 算法, 很好地展示了量子计算机的并行性。之后 Shor[5]在 1994 年提出了著名的 Shor 算法, 证明该算法可以在多项式时间完成大数因子分解问题。1996 年, Grover[6]在经典无序搜索算法的基础上提出了 Grover 算法, 该算法结合了幅度放大技术, 相较于经典算法实现了平方加速。近年来, 研究人员还发现可以利用量子计算高效地完成机器学习任务, 提出了一系列量子机器学习算法, 如量子线性回归[7, 8]、量子降维算法[9−12]、量子聚类[13−18]等。

作为机器学习的主要方法之一, 聚类分析常用于对未知类别的数据进行划分, 已广泛应用在销售、医学和生物等领域。在聚类分析中, 按照一定的规则将样本数据划分成若干个簇, 并把相似的样本聚在同一个簇中, 不相似的样本分在不同簇中。在 2013 年, Lloyd 等[19]提出了量子无监督学习, 指出由绝热算法实现的量子 *K*-means 算法可以在维数和样本数量参数上实现对经典 *K*-means 算法的指数加速。2019 年, Kerenidis 等[20]提出了 *q*-means 算法。与经典的 *K*-means 算法相比, 该算法提供了对数据数量的指数级加速。上述算法均需使用 QRAM 加载样本数据, 并且需要与数据量相当的存储空间。除此之外, QRAM 尚处于理论模型阶段[21], 在制备任意量子态方面是困难的。

本文结合角编码对数据进行加载, 基于已有样本对样本分量分别进行参数阈值设置, 执行编码数据、相似度度量、量子最小值搜索和质心迭代更新四个主要步骤。理论分析表明本文所提出算法相比于经典算法在运行时间上有平方级加速。

1 预备知识

1.1 经典 *K*-means 算法

K-means 算法是一种无监督的聚类算法。给定的样本集分成 *K* 个簇, 此算法将样本集中的每一个样本依次与 *K* 个簇的质心进行距离计算, 按照距离大小确定各个样本点最近质心的簇。经典 *K*-means 算法主要分为以下四个步骤: 1) 首先选取 *K* 个(*K* 可根据某个损失函数确定)质心, 通常是随机选取; 2) 计算余下的每一个样本点到各个质心的欧式距离, 并将其归入相互间距离最小的质心所在的簇; 3) 在所有样本点都划分完毕后, 重新计算各个簇的质心(通常是计算簇中样本点的均值), 然后迭代计算样本点到各个质心的距离, 并对所有样本点重新进行划分; 4) 重复第 2)、3) 步, 直到迭代计算后所有样本点的划分情况保持不变或小于误差, 此时 *K*-means 算法得到最优解, 将运行结果返回。

1.2 经典 *K*-means 算法相似度度量方法

经典 *K*-means 算法的关键步骤是计算待标记的样本到各个质心的距离, 并将其归入到二者间距离最小的质心所在的簇。其中, 各个簇的质心是通过计算簇中所有数据点的均值来确定的。常见的度量相似度的

方式有两种, 分别是用内积计算与欧式距离计算的结果来衡量相似度。考虑两个 N 维向量 $\boldsymbol{x}^{(i)}=\left(x_1^{(i)},x_2^{(i)},\cdots,x_N^{(i)}\right)$和$\boldsymbol{y}=\left(y_1,y_2,\cdots,y_N\right)$, 其基于内积的距离表达式可表示为 $\left|\boldsymbol{x}^j-\boldsymbol{y}\right|=\left|\boldsymbol{x}^j\right|\left|\boldsymbol{y}\right|-\boldsymbol{x}^j\boldsymbol{\cdot}\boldsymbol{y}$, 该距离主要关注两个向量之间的角度关系; 对于欧式距离, 其距离表达式为 $\left|\boldsymbol{x}^j-\boldsymbol{y}\right|=\sqrt{\sum_{i=1}^{N}\left(x_i^j-y_i\right)^2}$ 。这两种相似度度量方式的局限性在于特征尺度的差异将影响相似度度量。

量子 *K*-means 算法常采用欧氏距离作为度量距离的手段。该度量方式的局部较大特征将会降低较小数值特征的作用, 甚至使其根本不起作用。为了解决欧氏距离度量特征差异大的问题, 本研究在编码数据部分利用角编码对特征执行参数化预处理, 因此, 在相似度度量方面可有效避免因局部特征的数值太大而掩盖其他较小特征的情况。

1.3　本研究所提出算法涉及的量子门操作

经典计算机处理信息的基本单元是比特, 其状态为 0 或者 1。与此相似的是, 量子计算机处理的基本单元是 $|0\rangle$ 和 $|1\rangle$, 任意经过酉算子变化的单量子比特的状态可用二维希尔伯特空间里的一个单位复向量描述, 如

$$\left|\mu\right\rangle=\alpha_0|0\rangle+\alpha_1|1\rangle=\begin{bmatrix}\alpha_0\\ \alpha_1\end{bmatrix}, \tag{1}$$

式中: α_0 和 α_1 是复数, 且满足$\left|\alpha_0\right|^2+\left|\alpha_1\right|^2=1$。

量子逻辑门是作用于单个或多个量子比特以实现某个变换的酉算子操作。本研究需要用到的单量子比特逻辑门可表示为

$$\boldsymbol{H}=\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix},\ \boldsymbol{R}_y(\boldsymbol{\theta})=\begin{bmatrix}\cos\dfrac{\theta}{2} & -\sin\dfrac{\theta}{2}\\ \sin\dfrac{\theta}{2} & \cos\dfrac{\theta}{2}\end{bmatrix}, \tag{2}$$

其中 H 门主要用于创建叠加态, $R_y(\theta)$ 门主要用于旋转数据这一步骤。常见的还有双量子逻辑门, 与文中相联系的有 *SWAP* 门, 可表示为

$$\boldsymbol{SWAP}=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}. \tag{3}$$

构建本研究需要用到受控交换门$(C-SWAP)$门, 利用该量子门可以进行交换测试。不难看出, 控制量子比特为 $|1\rangle$ 时, 对目标比特作 *SWAP* 操作; 而处于 $|0\rangle$ 时, 目标量子态不变, 故而可得受控交换门为 $\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{SWAP}\end{bmatrix}$。需要注意的是, 矩阵中的元素都是 4×4 的矩阵。

最后介绍量子比较器(QCMP), 它一般被用作量子子程序, 此处将其作为一个算子考虑, 即

$$QCMP\left(|M\rangle\sum_{i=1}^{N}|i\rangle|0\rangle\right)=|M\rangle\left(\sqrt{\frac{R}{N}}\sum_{i<M}|i\rangle|1\rangle+\sqrt{\frac{N-R}{N}}\sum_{i\geq M}|i\rangle|0\rangle\right), \tag{4}$$

式中 $|M\rangle$ 表示预设需要比较的值, 在QCMP 作用后, 如果小于 $|M\rangle$ 会将辅助量子位设定为 $|1\rangle$, 否则为 $|0\rangle$, R 值为叠加态中小于 $|M\rangle$ 的数量。

2 量子算法描述

2.1 编码数据

数据角编码是制备量子态的重要方法, 可以有效地提高制备量子态的效率。在编码数据中, 本研究着重解决参数设置和数据加载这两个问题, 图 1 描述了由经典数据到量子线路转化的过程。

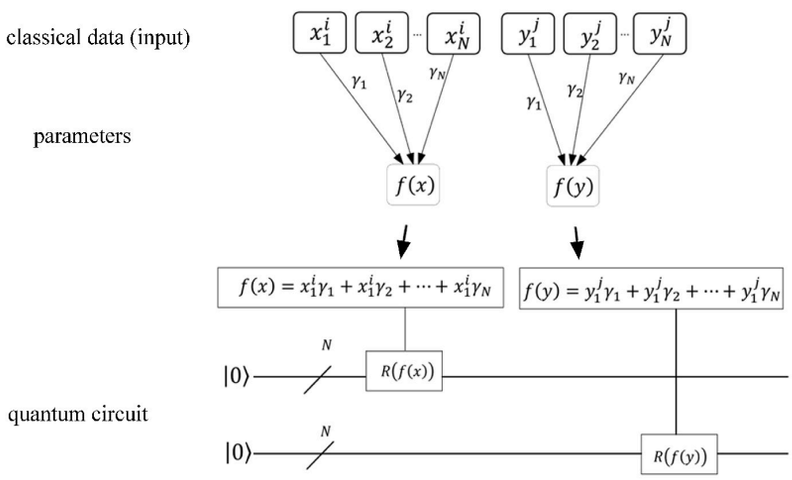

图 1 从经典数据到量子态的量子线路图

Fig. 1 Quantum circuit diagram from classical data to quantum state

首先考虑带标签的经典数据 $\boldsymbol{x}^{(i)}=\left(x_1^{(i)},x_2^{(i)},\cdots,x_N^{(i)}\right)$ 和带有标签的质心数据 $\boldsymbol{y}^{(j)}=\left(y_1^{(j)},y_2^{(j)},\cdots,y_N^{(j)}\right)$。对于所有的数据及其分量, 上标 i 表示样本标签, 且 $i\in\{1,2,\ldots,M\}$; j 表示簇标签, 且 $j\in\{1,2,\ldots,K\}$; 上标箭头表示经典数据; 下标表示分量的位置。

步骤 S1: 参数值的设定与分析

针对经典数据 $\boldsymbol{x}^{(i)}$ 的各个分量, 先设定各分量对应的参数 γ_i。基于编码的性质[22−24], 结合参数对数据进行编码时旋转角不超过周期 2π。超过 2π 将无法正确计算相似度, 即对应的用 Bloch 球表示的旋转角度过大会影响相似度的评估。其次需要对最终相似度结果的范围进行限制, 以约束参数的范围。因此, 先规定各分量数据与阈值化参数的关系应符合 $0\le\gamma_i y_i^{(j)},\gamma_i x_i^{(i)}\le 2\pi$, 有 $\gamma_i\le\dfrac{2\pi}{S_i}$, 其中 $S_i=\max\left(\max_{i,j}\left(x_i^{(i)}\right),y_i^{(j)}\right)$。对数据 $\boldsymbol{x}^{(i)}$ 和 $\boldsymbol{y}^{(j)}$ 的分量编码分别得到量子态

$$\begin{cases}\left|x_i^{(i)}\right\rangle=R_y\left(\gamma_i x_i^{(i)}\right)|0\rangle=\left[\cos\dfrac{\gamma_i}{2}x_i^{(i)}|0\rangle+\sin\dfrac{\gamma_i}{2}x_i^{(i)}|1\rangle\right]\\ \left|y_i^{(j)}\right\rangle=R_y\left(\gamma_i y_i^{(j)}\right)|0\rangle=\left[\cos\dfrac{\gamma_i}{2}y_i^{(j)}|0\rangle+\sin\dfrac{\gamma_i}{2}y_i^{(j)}|1\rangle\right]\end{cases}. \tag{5}$$

对 (5) 式的两个量子态作 $C-SWAP$ 操作, 便可得到任意分量 $x_i^{(i)}$ 和 $y_i^{(j)}$ 分量之间的保真度为 $\cos^2\dfrac{\gamma_i}{2}\left(x_i^{(i)}-y_i^{(j)}\right)$, 此处蕴含单调关系 $-\dfrac{\pi}{2}\le\dfrac{\gamma_i}{2}\left(x_i^{(i)}-y_i^{(j)}\right)\le\dfrac{\pi}{2}$, 即 $\gamma_i\le\dfrac{\pi}{\left|x_i^{(i)}-y_i^{(j)}\right|}$。因此, 先设定 $T_i=\max_{i,j}\left|x_i^{(i)}-y_i^{(j)}\right|$, 进一步令 $L_i=\max\left(S_i,2T_i\right)$, 可得表达式 $\gamma_i\equiv\dfrac{2\pi}{L_i}$。此处, 参数 γ_i 的设定最终确定特征缩放的范围, 解决了不同特征间

的尺度问题。最后，分别记编码质心和待分配样本的算子为 $U\left(\gamma;\boldsymbol{y}^{(j)}\right)\equiv\bigotimes_{i'=1}^{N}R_y\left(\gamma_{i'}y_{i'}^{(j)}\right)$ 和 $U\left(\gamma;\boldsymbol{x}^{(i)}\right)\equiv\bigotimes_{i'=1}^{N}R_y\left(\gamma_{i'}x_{i'}^{(i)}\right)$。

步骤S2: 经典数据的加载

利用酉算子 $U\left(\gamma;\boldsymbol{y}^{(j)}\right)$ 编码质心 $\boldsymbol{y}^{(j)}$, 得到量子态

$$\left|y^{(j)}\right\rangle=\left(\bigotimes_{i'=1}^{N}R_y\left(\gamma_{i'}y_{i'}^{(j)}\right)\right)|0\cdots0\rangle=\bigotimes_{i'=1}^{N}\left[\cos\frac{\gamma_{i'}}{2}y_{i'}^{(j)}|0\rangle+\sin\frac{\gamma_{i'}}{2}y_{i'}^{(j)}|1\rangle\right],\tag{6}$$

同理, 用相同的方式对向量 $\boldsymbol{x}^{(i)}$ 编码可以得到

$$\left|x^{(i)}\right\rangle=\left(\bigotimes_{i'=1}^{N}R_y\left(\gamma_{i'}x_{i'}^{(i)}\right)\right)|0\cdots0\rangle=\bigotimes_{i'=1}^{N}\left[\cos\frac{\gamma_{i'}}{2}x_{i'}^{(i)}|0\rangle+\sin\frac{\gamma_{i'}}{2}x_{i'}^{(i)}|1\rangle\right],\tag{7}$$

然后, 初始化量子寄存器1、2为 $|0\rangle_1^{\otimes\log_2K}$、$|0\rangle_2^{\otimes\log_2M}$, 用于构建条件受控的算子。与利用QRAM模型对数据的加载方式不同, 本研究以算子的方式将数据载入量子态中。再附加两个量子寄存器 $|0\rangle_3^{\otimes N}$、$|0\rangle_4^{\otimes N}$, 根据参考文献 [25] 对受控量子门的描述, 用量子寄存器1、2分别控制量子寄存器3、4对数据的载入。量子寄存器1控制量子寄存器3的算子形式为

$$\boldsymbol{U}_{\gamma\boldsymbol{y}}\equiv U_1\left(\gamma;\boldsymbol{y}^{(1)}\right)\oplus\cdots\oplus U_K\left(\gamma;\boldsymbol{y}^{(K)}\right)\equiv\begin{bmatrix}U_1\left(\gamma;\boldsymbol{y}^{(1)}\right)&\cdots&0\\\vdots&\ddots&\vdots\\0&\cdots&U_K\left(\gamma;\boldsymbol{y}^{(K)}\right)\end{bmatrix}_{\left(K\times2^N\right)\times\left(K\times2^N\right)}.\tag{8}$$

同理, 量子寄存器2控制量子寄存器4的算子 $\boldsymbol{U}_{\gamma\boldsymbol{x}}$ 与(8)式形式类似, 都是通过条件受控构建的。$\boldsymbol{U}_{\gamma\boldsymbol{y}}$ 和 $\boldsymbol{U}_{\gamma\boldsymbol{x}}$ 算子的构建依赖于叠加态 $\frac{1}{\sqrt{K}}\sum_{j=1}^{K}|j\rangle$、$\frac{1}{\sqrt{M}}\sum_{i=1}^{M}|i\rangle$ 的制备[26]。制备过程用到 $\boldsymbol{H}$ 门、受控酉门和量子比较器子程序。利用量子寄存器1控制编码算子 $U\left(\gamma;\boldsymbol{y}^{(j)}\right)$, 实现了对质心叠加态的制备。$\boldsymbol{U}_{\gamma\boldsymbol{y}}$ 对量子寄存器3的加载表现为

$$\boldsymbol{U}_{\gamma\boldsymbol{y}}\left(\frac{1}{\sqrt{K}}\sum_{j=1}^{K}|j\rangle_1|0\rangle_3^{\otimes N}\right)=\frac{1}{\sqrt{K}}\sum_{j=1}^{K}|j\rangle_1U_j\left(\gamma;\boldsymbol{y}^{(j)}\right)|0\rangle_3^{\otimes N}.\tag{9}$$

上述操作后，利用量子寄存器2控制编码算子 $U\left(\gamma;\boldsymbol{x}^{(i)}\right)$ 可得到状态 $\frac{1}{\sqrt{K}}\sum_{j=1}^{K}|j\rangle_1\frac{1}{\sqrt{M}}\sum_{j=1}^{M}|i\rangle_2\left(\bigotimes_{i'=1}^{N}\left(\cos\frac{\gamma_{i'}}{2}y_{i'}^{(j)}|0\rangle+\sin\frac{\gamma_{i'}}{2}y_{i'}^{(j)}|1\rangle\right)\right)_3\left(\bigotimes_{i'=1}^{N}\left[\cos\frac{\gamma_{i'}}{2}x_{i'}^{(i)}|0\rangle+\sin\frac{\gamma_{i'}}{2}x_{i'}^{(i)}|1\rangle\right]\right)_4$。为了便于分析，将其化简为 $\frac{1}{\sqrt{K}}\sum_{j=1}^{K}|j\rangle_1\frac{1}{\sqrt{M}}\sum_{j=1}^{M}|i\rangle_2\left|y^{(j)}\right\rangle_3\left|x^{(i)}\right\rangle_4$, 其中 $\left|y^{(j)}\right\rangle$ 表示量子寄存器3, $\left|x^{(i)}\right\rangle$ 表示量子寄存器4。

2.2　相似度度量

K-means 算法的一个关键步骤是估计测试样本与质心之间的相似度, 着重解决测试样本与质心之间距离的问题。为计算样本之间的相似度, 本研究采用了多量子比特交换测试线路[27]。

步骤S3: 多量子比特交换测试的实现

首先初始化长度为 l 的量子寄存器5, 利用 $H^{\otimes l}$ 门制备叠加态 $\frac{1}{\sqrt{2^l}}\sum_{q=0}^{2^l-1}|q\rangle$, 其中 $l=\lceil \log_2 N+1 \rceil$ 。对量子寄存器5中 q 的可能输出值做经典预处理, 具体过程如下: 1)对 q 值做取模运算, 以 x 值代替取模值, 即 $x=q \bmod 4$; 2) 将步骤1) 中的 x 值代入函数 $f(x)=\frac{x^2-3x+2}{2}$, 得到映射值0或1; 3) 将所有映射值为1的 q 值用二进制表示, 用来控制构建受控 *SWAP* 门。

以3量子比特为例构建多量子比特交换测试线路, 如图2所示。

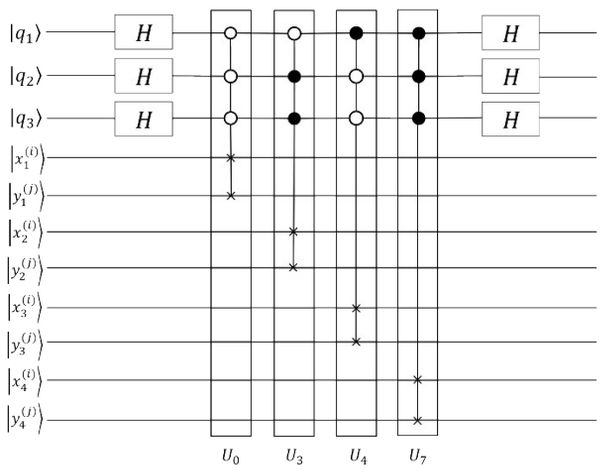

图2 3-辅助量子比特的线路示意图

Fig. 2 Circuit diagram with 3-auxiliary qubits

推广到 l 个控制量子比特, 则图2标记的算子数量为 2^{l-1}, 每一个算子可用

$$U_p=\begin{cases}\left(|4t\rangle\langle 4t|\otimes I_{4t}\otimes \boldsymbol{SWAP}_2+\sum_{i\neq 4t}|q\rangle\langle q|\otimes I_{4t+2}\right)\otimes I_{2N-4t-2}, & \text{if } p=4t\\ \left(|3+4t\rangle\langle 3+4t|\otimes I_{2+4t}\otimes \boldsymbol{SWAP}_2+\sum_{i\neq 4t+3}|q\rangle\langle q|\otimes I_{4+4t}\right)\otimes I_{2N-4t-4}, & \text{if } p=3+4t\\ I_{2N+l} & \text{otherwise}\end{cases} \tag{10}$$

表示, 其中算子下标表示量子比特数。由步骤S2可得量子态 $\frac{1}{\sqrt{K}}\sum_{j=1}^{K}|j\rangle_1\frac{1}{\sqrt{M}}\sum_{j=1}^{M}|i\rangle_2|y^{(j)}\rangle_3|x^{(i)}\rangle_4$, 在此基础上增加一个寄存器5, 初始化为 $|0\rangle^{\otimes l}$, 并执行多量子比特交换测试。先通过 $H^{\otimes l}$ 将寄存器5转换为叠加态, 得到量子态 $\frac{1}{\sqrt{K}}\sum_{j=1}^{K}|j\rangle_1\frac{1}{\sqrt{M}}\sum_{j=1}^{M}|i\rangle_2|y^{(j)}\rangle_3|x^{(i)}\rangle_4\left(\frac{1}{\sqrt{2^l}}\sum_{q=0}^{2^l-1}|q\rangle\right)_5$。下一步, 执行多量子比特交换测试以获得相似度度量的结果。为了方便表示, 记(10)式所有受控酉门的乘积形式为

$$U_\eta=U_{2N-1}U_{2N-2}\cdots U_1U_0, \tag{11}$$

式中算子 U_η 中有一半是多维单位阵。在执行完受控交换测试操作之后, 将得到量子态 $\frac{1}{\sqrt{K}}\sum_{j=1}^{K}|j\rangle_1\frac{1}{\sqrt{M}}\sum_{j=1}^{M}|i\rangle_2\left(I_{2N}\otimes H^{\otimes l}\right)U_\eta\left(|y^{(j)}\rangle_3|x^{(i)}\rangle_4\left(\frac{1}{\sqrt{2^l}}\sum_{q=0}^{2^l-1}|q\rangle\right)_5\right)$。测量量子比特 $|q_l\rangle$, 可以得到关于

$P\left(\left|q_l\right\rangle=\left|0\right\rangle\right)$的概率分布, 用来衡量质心 $\boldsymbol{y}^{(j)}$和样本 $\boldsymbol{x}^{(i)}$的相似度, 即

$$P\left(\left|q_l\right\rangle=\left|0\right\rangle\right)=\frac{\left[\frac{1}{2^{l-1}}\sum_{i'=0}^{2^{l-1}-1}\cos^2\ \frac{\gamma_{i'}}{2}\left(x_{i'}^{(i)}-y_{i'}^{(j)}\right)\right]+1}{2}=\frac{\left[\frac{1}{N}\sum_{i'=0}^{N-1}\cos^2\ \frac{\gamma_{i'}}{2}\left(x_{i'}^{(i)}-y_{i'}^{(j)}\right)\right]+1}{2},\tag{12}$$

式中 $N=2^{l-1}$ 表示维度。根据量子比特 $\left|q_l\right\rangle$ 的测量结果, 可以将相似度度量的结果改写为

$$\left|\psi_0\right\rangle=\frac{1}{\sqrt{K}}\sum_{j=1}^{K}\left|j\right\rangle_1\frac{1}{\sqrt{M}}\sum_{j=1}^{M}\left|i\right\rangle_2\left(\sin\theta_{ji}\left|u_{ji}\right\rangle\left|0\right\rangle+\cos\theta_{ji}\left|v_{ji}\right\rangle\left|1\right\rangle\right),\tag{13}$$

式中 $|0\rangle$、$|1\rangle$ 表示量子寄存器5第 l 量子位 q_l 的量子状态, $\left|u_{ji}\right\rangle$、$\left|v_{ji}\right\rangle$ 是两个复杂的量子态, 在分析过程中可忽略其具体形式。

步骤S4: 多量子相位估计

该步骤利用量子相位估计来获得所有 θ_{ji} 信息。为完成估计信息的任务, 还需要制备酉算子 $G=\prod_{ji}V_{ji}$, 过程如下:

1) 基于已定义的酉算子 U_η、 $U\left(\boldsymbol{\gamma};\boldsymbol{x}^{(i)}\right)$、$U\left(\gamma;\boldsymbol{y}^{(j)}\right)$来构建受控算子, 即对于所有标签$\left|ji\right\rangle$, 定义以下控制算子

$$\sum_{ji}V_{ji}\equiv\sum_{ji}\left|ji\right\rangle\left\langle ji\right|\ \otimes\left\{\left[I_{2N}\otimes\boldsymbol{H}^{\otimes l}\right]U_\eta\left[I_{2N}\otimes\boldsymbol{H}^{\otimes l}\right]\left[U\left(\gamma;\boldsymbol{x}^{(i)}\right)\otimes U\left(\gamma;\boldsymbol{y}^{(j)}\right)\otimes I_l\right]\right\}\left(\boldsymbol{H}^{\otimes\log_2K+\log_2M}\otimes I_{2N+l}\right).\tag{14}$$

2) 将控制算子$\sum_{ji}V_{ji}$作用于初始态$|0\rangle_{\log_2K+\log_2M}|0\rangle_{2N+l}$, 于是整个系统将处于

$$\left|\psi_0\right\rangle=\sum_{ji}V_{ji}\left|0\right\rangle_{\log_2K+\log_2M}\left|0\right\rangle_{2N+l}=\frac{1}{\sqrt{KM}}\sum_{ji}\left|ji\right\rangle\left(\sin\theta_{ji}\left|u_{ji}\right\rangle\left|0\right\rangle+\cos\theta_{ji}\left|v_{ji}\right\rangle\left|1\right\rangle\right)=$$
$$\frac{1}{\sqrt{KM}}\sum_{ji}\frac{-i}{\sqrt{2}}\left(\mathrm{e}^{\mathrm{i}\theta_{ji}}\left|\omega_{ji+}\right\rangle-\mathrm{e}^{-\mathrm{i}\theta_{ji}}\left|\omega_{ji-}\right\rangle\right),\tag{15}$$

式中$\left|\omega_{ji\pm}\right\rangle$ 表示 $\frac{1}{\sqrt{2}}\left|u_{ji}\right\rangle\left|0\right\rangle\pm\frac{\mathrm{i}}{\sqrt{2}}\left|v_{ji}\right\rangle\left|1\right\rangle$ 。

3) 由文献 [28], 可以构建迭代酉算子

$$G=\sum_{ji}V_{ji}\left(I^{\otimes\log_2K+\log_2M}\otimes\left(I^{\otimes2N+l}-2\left(\left|0\right\rangle^{\otimes(2N+l)}\left\langle0\right|\right)\right)\right)\sum_{ji}V_{ji}{}^{\dagger}\left(I^{\otimes\log_2K+\log_2M+2N+l-1}\otimes Z_{q_l}\right),\tag{16}$$

使得G作用到状态 $\left|\psi_0\right\rangle$, 将产生新的的量子态

$$\left|\psi_1\right\rangle=G\left|\psi_0\right\rangle=\frac{1}{\sqrt{KM}}\sum_{ji}\left|j\right\rangle\left|i\right\rangle\left(\sin3\theta_{ji}\left|u_{ji}\right\rangle\left|0\right\rangle+\cos3\theta_{ji}\left|v_{ji}\right\rangle\left|1\right\rangle\right).\tag{17}$$

因此, 在给定量子态 $\left|\psi_0\right\rangle$ 以及迭代酉算子$G=\prod_{i,j}V_{ji}$的条件下, 下一步可以执行多量子相位估计, 并将所有 θ_{ji} 的信息转到新的量子寄存器中。首先附加精度为 t 的量子寄存器6, 利用 G 做量子相位估计[29], 并对算法的结果进行分析得到量子态

$$\left|\psi_2\right\rangle=\frac{-i}{\sqrt{2MK}}\sum_{j=1}^{K}\sum_{i=1}^{M}\left[\mathrm{e}^{i\theta_{ji}}\left|2^t\frac{\theta_{ji}}{\pi}\right\rangle\left|ji\right\rangle\left|\omega_{ji+}\right\rangle-\mathrm{e}^{-i\theta_{ji}}\left|2^t\left(1-\frac{\theta_{ji}}{\pi}\right)\right\rangle\left|ji\right\rangle\left|\omega_{ji-}\right\rangle\right],\tag{18}$$

此式表示任意标签 ji 分别对应两种输出结果, 即 $\left|2^t\frac{\theta_{ji}}{\pi}\right\rangle$ 和 $\left|2^t\left(1-\frac{\theta_{ji}}{\pi}\right)\right\rangle$ 。多量子相位估计线路如图3所

示, 其中 QFT† 表示逆傅里叶变换。最后可在量子寄存器6得到二进制序列, 用于表示样本 i 和质心 j 的相似度信息。

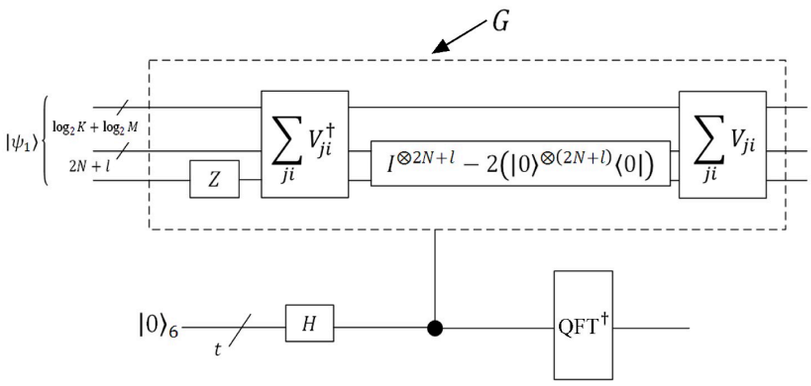

图3 多量子相位估计线路图

Fig. 3 Circuit diagram of multi-quantum phase estimation

步骤S5: 转移相位估计信息

根据(12)式保真度公式的结果 $\frac{1}{N}\sum_{i'=0}^{N-1}\cos^2\frac{\gamma_{i'}}{2}\left(x_{i'}^{(i)}-y_{i'}^{(j)}\right)=1-2\cos^2\left(\theta_{ji}\right)$, 以及条件 $\theta_{ji}\in\left[\frac{\pi}{4},\frac{3\pi}{4}\right]$, $\left|\cos\theta_{ji}\right|$ 愈大则保真度愈小。下一步是实现量子寄存器6信息的转化, 首先附加量子寄存器7, 然后利用量子寄存器6控制量子寄存器7从而得到相位信息的转化结果。为了保证在同一个标签上输出的信息是一致的, 利用量子线路构建 $\left|\cos\theta_{ji}\right|$ 函数。当输入是 $\frac{\theta_{ji}}{\pi}$ 和 $1-\frac{\theta_{ji}}{\pi}$ 时, 输出值对应 $\left|\cos\theta_{ji}\right|$。下一步构建量子余弦函数线路[30], 利用量子线路并行求解 $\left|\cos\theta_{ji}\right|$, 得到量子态

$$\frac{-i}{\sqrt{2KM}}\sum_{j=1}^{K}\sum_{i=1}^{M}\left[\mathrm{e}^{\mathrm{i}\theta_{ji}}\left|2^t\frac{\theta_{ji}}{\pi}\right\rangle|ji\rangle|\omega_{ji+}\rangle-\mathrm{e}^{-\mathrm{i}\theta_{ji}}\left|2^t\left(1-\frac{\theta_{ji}}{\pi}\right)\right\rangle|ji\rangle|\omega_{ji-}\rangle\right]\left|\left|\cos\theta_{ji}\right|\right\rangle_7,$$

其中 $\left|\left|\cos\theta_{ji}\right|\right\rangle$ 是数值 $\left|\cos\theta_{ji}\right|$ 二进制的多量子表示形式。由此, 成功创建了一个存储质心与样本之间相似度信息的量子叠加态。

2.3 量子最小值搜索

为方便起见, 将2.2节最终得到的量子态简化为 $\frac{1}{\sqrt{KM}}\sum_{j=1}^{K}|j\rangle_1\sum_{i=1}^{M}|i\rangle_2\left|\left|\cos\theta_j\right|\right\rangle_7$, 本节将利用量子最小值搜索算法[31]对其进行步骤描述。

步骤S6: 量子最小值搜索算法求最小值标签

1) 随机初始化一个标签, 确定其量子寄存器1、2的大小。针对量子态 $\frac{1}{\sqrt{KM}}\sum_{j=1}^{K}|j\rangle_1\sum_{i=1}^{M}|i\rangle_2\left|\left|\cos\theta_{ji}\right|\right\rangle_7$, 随机初始化一个阈值标签 ji, 并将对应量子寄存器7的值设定为 y。附加一个寄存器, 记作 $\frac{1}{\sqrt{KM}}\sum_{j=1}^{K}|j\rangle_1\sum_{i=1}^{M}|i\rangle_2\left|\left|\cos\theta_{ji}\right|\right\rangle_7|y\rangle_8$。

2) 利用 QCMP[32]作用于量子寄存器7和8。附加一个量子寄存器9存储标记信息, 得到

$\frac{1}{\sqrt{KM}}\left(\sum_{j\in G}|j\rangle_1\sum_{i=1}^{M}|i\rangle_2\left|\cos\theta_{ji}\right|\rangle_7|y\rangle_8|1\rangle_9+\sum_{j\notin G}|j\rangle_1\sum_{i=1}^{M}|i\rangle_2\left|\cos\theta_{ji}\right|\rangle_7|y\rangle_8|0\rangle_9\right)$, 其中 G 表示 $\left|\cos\theta_{ji}\right|\rangle$ 小于阈值 $|y\rangle$ 的标签集合, 此时将量子寄存器 9 置为 $|1\rangle$, 否则置为 $|0\rangle$。

3) 利用 Qsearch 算法[33]搜索量子寄存器 9 是否处于 $|1\rangle$。如果处于 $|0\rangle$, 则直接输出量子寄存器 8 对应的结果;否则读取量子寄存器 7 的信息, 并将该信息赋值到量子寄存器 8, 以重新确定新一轮循环的状态。

4) 根据文献[31]的结论, 当总时间复杂度小于 $O\left(22.5\sqrt{KM}+1.4\log_2^2 KM\right)$, 重复步骤 2) 和 3); 否则直接读取索引。

利用上述步骤可以得到待分配样本点所归属的簇标签。

2.4　质心迭代更新

将所有样本点分配到簇的过程会影响数据点的分布, 进而改变质心的位置, 因此, 在完成一轮迭代后需要重新考虑聚类效果。在聚类过程, 若样本点均匀分布在质心周围, 不影响质心分布的稳定性; 若样本点非均匀分布在质心周围, 则需要重新计算质心, 并对所有样本重新聚类。考虑到数据集在一轮迭代中可能改变原质心位置的情况, 就需要重新计算簇的质心。若需要对簇中样本计算新的质心, 这必然会导致质心计算的复杂度增加。为解决质心计算的问题引入了随机采样方案, 此方案利用量子的概率性输出簇中样本子集并近似代表质心, 可以降低计算质心的复杂度。

对每个簇进行随机采样之前, 记原始的簇大小为 $|C^j|$, 任一经过随机采样的簇的大小记为 $\left|\widetilde{C^j}\right|$, 满足关系式 $\left|\widetilde{C^j}\right|\le|C^j|$。质心迭代涉及到对所有新质心的计算, 并用随机采样的样本均值表示新质心。在计算得到新质心后, 需要重新按照 2.1～2.4 中的步骤对所有样本进行聚类操作。

3　性能分析

3.1　复杂度分析

在数据编码部分, 量子寄存器 1、2 创建叠加态的时间复杂度为 $O\left(\log_2 K+\log_2 M\right)$, 故可将整个编码数据部分的时间复杂度记为 $O\left(\log_2 KM\right)$。在步骤 S3 中, 其时间复杂度集中在 $U_\eta=U_{2N-1}U_{2N-2}\ldots U_1U_0$, 故为 $O(N)$。针对多量子相位估计步骤, 酉算子 G 对应的时间复杂度为 $O\left(\log_2 KM\left[\log_2 KM+N\right]\right)$。另外, 相位估计的复杂度还和量子寄存器 6 的精度 t 相关, 这意味着当精度 t 远小于样本维数 N 或者 $t\ll\log_2 KM$, 整个多量子相位估计的时间复杂度为 $O\left(\log_2 KM\left[\log_2 KM+N\right]\right)$。下一步来分析量子绝对值余弦函数 $\left|\cos\theta_{ji}\right|$, 对函数进行模块化处理, 其时间复杂度与数据维度及训练集大小无关, 所以时间复杂度记为 $O(1)$。2.3 节分析了量子最小值搜索的算法过程, 并给出执行该步骤的时间复杂度为 $O\left(22.5\sqrt{KM}+1.4\log_2^2(KM)\right)$。所以, 样本分配到簇的时间复杂度记为 $O\left(\log_2 KM\left[\log_2 KM+N\right]\left[22.5\sqrt{KM}+1.4\log_2^2(KM)\right]\right)$, 并且可以进一步简化为 $O(\log_2 KM\left[\log_2 KM+N\right]\sqrt{KM})$。

经典 *K*-means 算法、本算法和其他量子 *K*-means 算法时间复杂度对比如表 1 所示, 表中时间复杂度表示一轮迭代的总时间复杂度, 其中 M 表示样本数、η 是样本最大平方范数、δ 是误差参数。需要注意的是,

在经典 *K*-means 算法中还涉及对各个簇质心的计算。Lloyd 的算法实现了对维数和样本数的指数加速, Kerenidis 的算法实现了对样本数的指数加速, 二者算法的加速效果都是基于 QRAM 模型实现的。

表 1 *K*−means 算法之间的比较

Table 1 Comparison between *K*−means algorithms

Algorithm	Implementation mode	Time complexity
Classical *K*-means algorithm	/	$O(KMN)$
The proposed algorithm	Angle encoding	$O\left(\left(N+poly\left(\log_2 KM\right)\right)\sqrt{KM}\right)$
Lloyd'algorithm[19]	QRAM	$O\left(\log_2 KMN\right)$
Kerenidis'algorithm[20]	QRAM	$O\left(\frac{K^2N\eta^{2.5}}{\delta^3}poly\left(\log M\right)\right)$

3.2 数值实验

为了更加直观地判断样本的归属, 以二维样本 $\boldsymbol{y}$ (3, 30) 和质心 $\boldsymbol{x}^1$(8, 70)、 $\boldsymbol{x}^2$(2, 25) 为例进行数值实验。按照经典方式计算经预处理的数据, 样本点 $\boldsymbol{y}$ 与 质心点 $\boldsymbol{x}^2$ 的相似度更高。图 4 是该例在 qasm 量子模拟器上运行的实验结果分布图。针对低维样本数据的相似度度量任务, 由条形图的概率分布可计算出相似度度量结果。

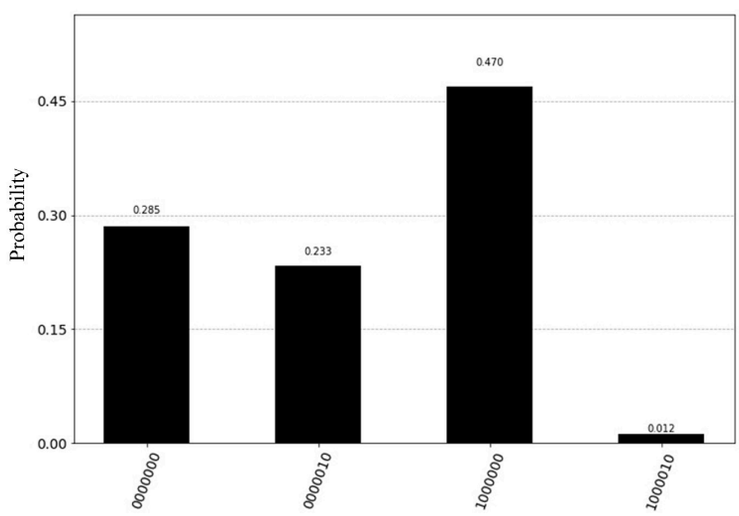

图 4 二维数据点到两个质心的概率分布

Fig. 4 Probability distribution of two-dimensional data points to two centroids

由图 4 可见, 第 7 量子比特表示对应标签所取得概率是大致相等的, 各接近 50%。在测得量子寄存器 7 结果为 1 情况下, 发现分别对第 2 个量子寄存器测得 0/1 的概率波动较大, 这意味着样本 $\boldsymbol{y}$ 和质心 $\boldsymbol{x}^1$ 之间的相似度较低, 使得二者测量概率接近; 在测得量子寄存器 7 为 0 的条件下, 所得第 2 量子寄存器为 0 的概率 $P(0)$ 远大于其概率为 1 的概率 $P(1)$, 这表明样本 $\boldsymbol{y}$ 和质心 $\boldsymbol{x}^2$ 非常接近, 使得测量的概率差别较大, 由(12)式还可以计算出样本和质心相似度的大小。计算可得: 样本 $\boldsymbol{y}$ 和质心 $\boldsymbol{x}^2$ 相似度为95.02%, 而和质心 $\boldsymbol{x}^1$ 的相似度为 10.04%。因此, 本研究所涉及线路可用于计算样本 $\boldsymbol{y}$ 与质心 $\boldsymbol{x}^2$ 的相似度, 并且该结果与经典计算得到的预测结果一致。

4 结　论

提出了一种无需 QRAM 存储的量子 *K*-means 算法。该算法利用角编码技术将经典数据转化为量子态, 并且对输入的经典数据施加不同参数, 从而解决样本不同特征尺度差异的问题。在相似度度量步骤, 使用多量子比特交换测试及量子相位估计算法, 以估计样本与质心之间的相似度信息; 在量子最小值搜索阶段, 将量子最小值搜索算法用于求解待分配样本点所归属的簇标签; 最后, 通过概率性输出样本子集近似代表质心。时间复杂度分析结果表明, 所提出算法相较于经典 *K*-means 算法实现了样本数的平方加速。还利用角编码加载了已预处理的数据, 由数值实验得出的相似度结果与经典结果一致。此外, 虽然本算法无法达到其他量子 *K*-means 算法的指数级加速效果, 但其可有效实现特征权重不同的数据集的聚类任务, 具有更广泛的适用范围。除了特征权重不同, 数据集的分布情况也是影响聚类效果的另一个主要因素。例如, 现有量子 *K*-means 算法对非凸数据集无法进行有效聚类分析。因此, 如何设计高效的量子 *K*-means 算法来解决非典型数据分布的数据集聚类问题将是下一步研究的重点。

参考文献:

[1] Feynman R P. Simulating physics with computers [J]. *International Journal of Theoretical Physic*, 1982, 21(6): 467-488.

[2] Lloyd S, Mohseni M, Rebentrost P. Quantum principal component analysis [J]. *Nature Physics*, 2014, 10(9): 631-633.

[3] Deutsch D. Quantum theory, the Church–Turing principle and the universal quantum computer [C]. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 1985: 97-117.

[4] Deutsch D E. Quantum computational networks [C]. *Proceedings of The Royal Society of London. A. Mathematical and Physical Sciences*, 1989: 73-90.

[5] Shor P W. Algorithms for quantum computation: discrete logarithms and factoring [C]. *Proceedings 35th Annual Symposium on Foundations of Computer Science*. Santa Fe, USA, IEEE, 1994: 124-134.

[6] Grover L K. A fast quantum mechanical algorithm for database search [C]. *Proceedings of The Twenty-eighth Annual ACM Symposium on Theory of Computing*, Philadelphia, USA, ACM, 1996: 212-219.

[7] Zhang D B, Xue Z Y, Zhu S L, *et al*. Realizing quantum linear regression with auxiliary qumodes [J]. *Physical Review A*, 2019, 99(1): 012331.

[8] Gilyén A, Song Z, Tang E. An improved quantum-inspired algorithm for linear regression [OL]. arXiv: 2009.07268, 2020, https://arxiv.org/abs/2009.07268.

[9] Sornsaeng A, Dangniam N, Palittapongarnpim P, *et al*. Quantum diffusion map for nonlinear dimensionality reduction [J]. *Physical Review A*, 2021, 104(5): 052410.

[10] Duan B J, Yuan J B, Xu J, *et al*. Quantum algorithm and quantum circuit for A-optimal projection: Dimensionality reduction [J]. *Physical Review A*, 2019, 99(3): 032311.

[11] Lin J, Bao W S, Zhang S, *et al*. An improved quantum principal component analysis algorithm based on the quantum singular threshold method [J]. *Physics Letters A*, 2019, 383(24): 2862-2868.

[12] He C, Li J Z, Liu W Q, *et al*. A low-complexity quantum principal component analysis algorithm [J]. *IEEE Transactions on Quantum Engineering*, 2022, 3: 1-13.

[13] Chen M H, Guo G D, Lin S. Quantum recommendation algorithm based on Hamming distance [J]. *Chinese Journal of Quantum Electronics*, 2021, 38(3): 332-340.
陈梦涵, 郭躬德, 林 崧. 基于汉明距离的量子推荐算法 [J]. 量子电子学报, 2021, 38(3): 332-340.

[14] Fan D C, Song Z L, Jon S, *et al*. An improved quantum clustering algorithm with weighted distance based on PSO and research on the prediction of electrical power demand [J]. *Journal of Intelligent & Fuzzy Systems*, 2020, 38(2): 2359-2367.

[15] Yu K, Guo G D, Li J, *et al*. Quantum algorithms for similarity measurement based on Euclidean distance [J]. *International Journal of Theoretical Physics*, 2020, 59(10): 3134-3144.

[16] Gong C Q, Dong Z Y, Gani A, *et al*. Quantum *K*-means algorithm based on trusted server in quantum cloud computing [J]. *Quantum Information Processing*, 2021, 20(4): 1-22.

[17] Wu Z H, Song T T, Zhang Y B. Quantum *K*-means algorithm based on Manhattan distance [J]. *Quantum Information Processing*, 2022, 21(1): 19.

[18] Khan S U, Awan A J, Vall-Llosera G. *K*-means clustering on noisy intermediate scale quantum computers [OL]. arXiv: 1909.12183, 2019, https://arxiv.org/abs/1909.12183.

[19] Lloyd S, Mohseni M, Rebentrost P. Quantum algorithms for supervised and unsupervised machine learning [OL]. arXiv*:* 1307. 0411, 2013, https://arxiv.org/abs/1307.0411.

[20] Kerenidis I, Landman J, Luongo A, *et al*. *q*-means: A quantum algorithm for unsupervised machine learning [C]. *Proceedings of the 32nd Advances in Neural Information Processing Systems*, Montreal, Canada, 2019: 4136–4146.

[21] Huang Y M, Lei H, Li X Y. A survey on quantum machine learning [J]. *Chinese Journal of Computers*, 2018, 41(1): 145-163.
黄一鸣, 雷 航, 李晓瑜. 量子机器学习算法综述 [J]. 计算机学报, 2018, 41(1): 145-163.

[22] Zang Y M, Zhu S C, Wei Z H, *et al*. A pseudo color coding method for quantum image [J]. *Chinese Journal of Quantum Electronics*, 2022, 39(3): 343-353.
臧一鸣, 朱尚超, 魏战红, 等. 一种量子图像伪彩色编码方法 [J]. 量子电子学报, 2022, 39(3): 343-353.

[23] Weigold M, Barzen J, Leymann F, *et al*. Expanding data encoding patterns for quantum algorithms [C]. *2021 IEEE 18th International Conference on Software Architecture Companion (ICSA-C)*, Stuttgart, Germany, 2021: 95-101.

[24] Schuld M. Supervised quantum machine learning models are kernel methods [OL]. 2021, arXiv: 2101.11020, https://arxiv.org/abs/2101.11020.

[25] Williams C P. *Explorations in Quantum Computing* [M]. 2nd ed., New York: Springer, 2011: 83-91.

[26] Dang Y J, Jiang N, Hu H, *et al*. Image classification based on quantum *K*-Nearest-Neighbor algorithm [J]. *Quantum Information Processing*, 2018, 17(9): 239.

[27] Li P C, Guo J H, Wang B, *et al*. Quantum circuits for calculating the squared sum of the inner product of quantum states and its application [J]. *International Journal of Quantum Information*, 2019, 17(5): 1950043.

[28] Zhao J, Zhang Y H, Shao C P, *et al*. Building quantum neural networks based on a swap test [J]. *Physical Review A*, 2019, 100 (1): 012334.

[29] Li P, Wang B. Quantum neural networks model based on swap test and phase estimation [J]. *Neural Networks*, 2020, 130: 152-164.

[30] Wang S B, Wang Z M, Li W D, *et al*. Quantum circuits design for evaluating transcendental functions based on a function-value binary expansion method [J]. *Quantum Information Processing*, 2020, 19(10): 347.

[31] Quek Y, Canonne C, Rebentrost P. Robust quantum minimum finding with an application to hypothesis selection [OL]. 2020, arXiv: 2003.11777, https://arxiv.org/abs/2003.11777.

[32] Xia H Y, Li H S, Zhang H, *et al*. An efficient design of reversible multi-bit quantum comparator via only a single ancillary bit [J]. *International Journal of Theoretical Physics*, 2018, 57(12): 3727-3744.

[33] Brassard G, Høyer P, Mosca M, *et al*. Quantum amplitude amplification and estimation [J]. *Contemporary Mathematics*, 2002, 305: 53-74.