doi:10.3969/j.issn.1001-4616.2021.04.015

基于量子计数的贝叶斯二元分类算法

陆春悦,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 350117)

[摘要]　贝叶斯分类算法是一种基于概率统计理论的有监督学习算法,常被用于分类问题中. 本文将量子计数与经典贝叶斯分类算法相结合,提出一种新的量子贝叶斯分类算法. 通过量子随机访问存储器制备所需的量子态,使用 oracle 进行相位翻转并构造与之所对应的操作算子,在操作算子的本征态空间上重新描述量子态,借助辅助粒子进行相位估计,投影测量后即可高效地计算出贝叶斯分类所需的数据,实现量子贝叶斯分类算法. 该算法在低维特征空间中与经典算法相比有着指数级加速.

[关键词]　量子机器学习,贝叶斯分类,二元分类,量子计数,相位估计
[中图分类号]TP38;TP181　[文献标志码]A　[文章编号]1001-4616(2021)04-0117-05

Bayesian Binary Classification Algorithm Based on Quantum Counting

Lu Chunyue, Guo Gongde, Lin Song

(College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China)

Abstract: Bayesian classification algorithm is a supervised learning algorithm based on the statistics theory of probability, which is often used in classification problems. In this paper, a new quantum Bayesian classification algorithm is proposed by combining quantum counting with classical Bayesian classification algorithm. The required quantum states are prepared by a quantum random access memory, the oracle is used to phase flip and construct the corresponding operator, the quantum states are redescribed on the eigenstate space of the operator, and phase estimation is performed with the help of auxiliary particles. Then, the data required for Bayesian classification can be efficiently calculated after projection measurements and the quantum Bayesian classification algorithm can be realized. Compared with the classical algorithm, this algorithm has exponential acceleration in the low dimensional feature space.

Key words: quantum machine learning, Bayesian classification, binary classification, quantum counting, phase estimation

量子计算利用量子并行、量子纠缠等特性解决一些计算任务,展现出了比经典计算更为优越的计算能力[1]. 1994 年,Shor[2] 提出用于大数分解的 Shor 算法,与经典算法相比,实现了指数级加速. 1997 年,Grover[3] 提出了针对非结构化数据库的量子搜索算法,时间复杂度为 $O\sqrt{N}$,达到了二次加速. 受这些量子算法的启发,人们对量子机器学习进行研究,提出了一系列高效的量子机器学习算法,如量子 K 近邻算法[4-5]、量子决策树[6]、量子支持向量机[7-8]、量子关联规则[9-10]等.

贝叶斯分类算法是一种常见的机器学习算法,它利用贝叶斯定理与联合概率模型进行分类预测,被广泛应用于文本分类. Shao[11] 在 2020 年提出了基于块编码的量子贝叶斯分类算法(简记为 Shao 算法). 该算法将块编码与贝叶斯分类相结合,实现了指数级加速. 然而,该算法仅仅适用于厄米矩阵. 本文针对这一问题进行研究,提出了一种基于量子计数的贝叶斯二元分类算法. 该算法通过量子计数与相位估计,快速得到能够反映待分类数据属于第 k 类别概率的相关值,获取待分类数据所属类别. 本文所提算法在低维特征空间中与经典算法相比有着指数级加速,也可应用于更为普遍的数据集.

收稿日期:2021-07-12.
基金项目:国家自然科学基金项目(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项目(2018J01776).
通讯作者:林崧,博士,教授,博士生导师,研究方向:量子机器学习. E-mail:lins95@fjnu.edu.cn

1　背景知识

1.1　朴素贝叶斯分类算法

朴素贝叶斯分类算法是在特征条件独立假设和贝叶斯定理的基础上得出的一种分类算法,它利用贝叶斯公式可以得到待分类数据属于不同类别的概率[12-14]. 在该问题中,本文假设存在一个由 N 个有类标签的数据$(\boldsymbol{x}_i,\boldsymbol{y}_i)$ $(i=1,2,\cdots,N)$构成的数据集 T,这里,$\boldsymbol{x}_i^{\mathrm{T}}=(\boldsymbol{x}_i^{(1)},\boldsymbol{x}_i^{(2)},\cdots,\boldsymbol{x}_i^{(M)})^{\mathrm{T}}$ 表示有 M 个属性的向量,$\boldsymbol{y}_i$ 表示 $\boldsymbol{x}_i$ 的类别,其中 $\boldsymbol{y}_i\in\boldsymbol{c}_k$,$k=1,2,\cdots,K$. 更多地,$\boldsymbol{x}_i^{(j)}\in\{\boldsymbol{a}_{jl}\}_{l=1}^{s_j}$表示 $\boldsymbol{x}_i$ 的第 j 个属性,其中 s_j 表示第 j 个属性种类. 贝叶斯分类的目的是判断新测试样本数据 $\tilde{\boldsymbol{x}}$ 所属的类别 $\tilde{\boldsymbol{y}}$,在该模型下,算法是先计算出新数据 $\tilde{\boldsymbol{x}}$ 属于类别 $\boldsymbol{c}_k$ 的概率 $P(\boldsymbol{c}_k|\tilde{\boldsymbol{x}})$,然后判断出其最大可能属于的类别.

根据贝叶斯定理和联合概率分布,$P(\boldsymbol{c}_k|\tilde{\boldsymbol{x}})$可以表述为:

$$P(\boldsymbol{c}_k|\tilde{\boldsymbol{x}})=\frac{P(\boldsymbol{X}=\tilde{\boldsymbol{x}},\boldsymbol{Y}=\boldsymbol{c}_k)}{P(\tilde{\boldsymbol{x}})}=\frac{P(\boldsymbol{y}_i=\boldsymbol{c}_k)\prod_{j=1}^{M}P(\boldsymbol{x}_i^{(j)}=\tilde{\boldsymbol{x}}_j\mid\boldsymbol{y}_i=\boldsymbol{c}_k)}{P(\tilde{\boldsymbol{x}})}. \tag{1}$$

通过式(1),容易发现对于待分类数据 $\tilde{\boldsymbol{x}}$,仅需要通过比较 $P(\tilde{\boldsymbol{x}},\boldsymbol{y}_i=\boldsymbol{c}_k)$就可得到 $\tilde{\boldsymbol{x}}$ 所属类别,

$$\tilde{\boldsymbol{y}}=\arg\max_{\boldsymbol{c}_k}P(\boldsymbol{Y}_i=\boldsymbol{c}_k)\prod_{j=1}^{M}P(\boldsymbol{X}_i^{(j)}=\tilde{\boldsymbol{x}}_j|\boldsymbol{Y}_i=\boldsymbol{c}_k). \tag{2}$$

因此,贝叶斯分类过程要求计算概率,

$$P(\boldsymbol{Y}_i=\boldsymbol{c}_k)=\frac{\sum_{i=1}^{N}I(\boldsymbol{y}_i=\boldsymbol{c}_k)}{N},k=1,2,\cdots K, \tag{3}$$

式中,$\sum_{i=1}^{N}I(\boldsymbol{y}_i=\boldsymbol{c}_k)$表示在数据集 T 中 $\boldsymbol{y}_i=\boldsymbol{c}_k$ 的个数. 此外,还需计算,

$$P(\boldsymbol{X}=\tilde{\boldsymbol{x}}|\boldsymbol{Y}_i=\boldsymbol{c}_k)=\prod_{j=1}^{M}P(\boldsymbol{X}_i^{(j)}=\tilde{\boldsymbol{x}}_j|\boldsymbol{Y}_i=\boldsymbol{c}_k). \tag{4}$$

显然,后者是算法的主要步骤. 这样,经典贝叶斯分类算法所需的时间复杂度为 $O(NM^2+M^2)$. 因此,如何高效地执行该步骤,是提高整个算法效率的关键. 在第 2 节中,本文为该步骤设计了相应的量子算法. 与经典贝叶斯二元分类算法相比,该量子算法能够指数级地降低时间复杂度.

1.2　量子计数

量子计数是相位估计在 Grover 迭代算子 G 上的应用[15]. 它通过估计 G 的特征值,来估计一个未知 Q 元搜索问题中解的个数$|\psi_1\rangle=\frac{\sum_{f(q)=1}|q\rangle}{\sqrt{D}}$. 主要思路如下:在一个 Q 元的无序数据库中,计算出问题解的个数 D. 搜索问题空间由量子态$|\psi\rangle$来表示,

$$|\psi\rangle=\sqrt{\frac{D}{Q}}\frac{\sum_{f(q)=1}|q\rangle}{\sqrt{D}}+\sqrt{\frac{Q-D}{Q}}\frac{\sum_{f(q)=0}|q\rangle}{\sqrt{Q-D}}=\sin(\theta)|\psi_1\rangle+\cos(\theta)|\psi_0\rangle, \tag{5}$$

式中,$|\psi_1\rangle=\frac{\sum_{f(q)=1}|q\rangle}{\sqrt{D}}$ 表示搜索问题解的和,$f(q)=1$ 表示$|q\rangle$对应的量子态是问题的一个解;$|\psi_0\rangle=\frac{\sum_{f(q)=0}|q\rangle}{\sqrt{Q-D}}$ 代表非问题解的和. 由于 $\sin^2\theta=\frac{D}{Q}$,且我们已经知道问题空间的规模为 Q,所以可以利用 θ 来估计解的个数 D.

更多地,此无序搜索问题中的 Grover 迭代算子 $G=(I-2|\psi\rangle\langle\psi|)(I-2|\psi_1\rangle\langle\psi_1|)$拥有本征值 $\lambda_{\pm}=\mathrm{e}^{\pm2\mathrm{i}\theta}$ $(\mathrm{i}=\sqrt{-1})$及相对应的本征态$|\psi_{\pm}\rangle=\frac{|\psi_0\rangle\mp\mathrm{i}|\psi_1\rangle}{\sqrt{2}}$. 那么,在 G 算子的本征态空间上重新描述$|\psi\rangle$,可得:

$$|\psi\rangle=\frac{\mathrm{e}^{\mathrm{i}\theta}|\psi_+\rangle+\mathrm{e}^{-\mathrm{i}\theta}|\psi_-\rangle}{\sqrt{2}}. \tag{6}$$

进一步,对 G 算子进行相位估计可得整个系统空间量子态为:

$$|\varphi\rangle=\frac{\mathrm{e}^{\mathrm{i}\theta}|2\theta\rangle|\psi_+\rangle+\mathrm{e}^{-\mathrm{i}\theta}|2\pi-2\theta\rangle|\psi_-\rangle}{\sqrt{2}}. \tag{7}$$

然后,在计算机上对寄存器$|2\theta\rangle$进行测量,可以获得 θ 的估计值,从而得到问题的解的个数 D.

2 基于量子计数的贝叶斯二元分类算法

本节所提出的量子贝叶斯二元分类算法,主要分为 4 个步骤:制备量子初态;量子计数;量子测量得到相关概率;通过后续简单的计算,确定测试样本的类别. 整个量子算法电路图如图 1 所示.

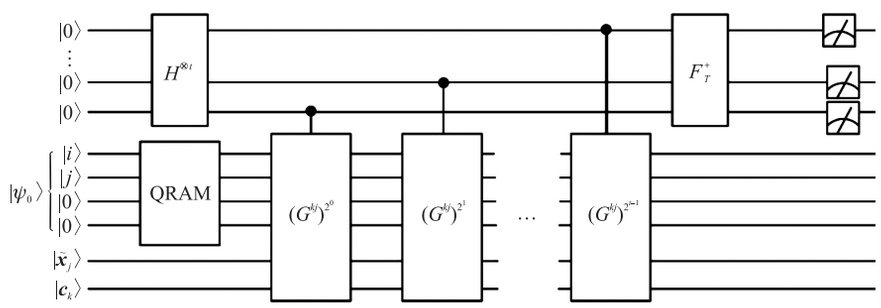

图 1　量子贝叶斯二元分类算法的整体量子电路图

Fig. 1　The overall quantum circuit diagram of quantum Bayesian binary classification algorithm

步骤 1:制备量子初态

考虑 M 维向量,将该经典数据映射到量子态,需要 $\log_2 M$ 个量子比特来存储该数据. 这里,利用量子随机访问存储器[16](quantum random access memory,QRAM)来并行访问数据,并以相干量子叠加方式进行内存访问. 假设 R 是一个地址寄存器,它包含叠加的地址(为叠加地址的振幅),QRAM 将返回 1 个属于该地址寄存器的叠加量子态. 如式(8)所示:

$$\sum_\alpha \phi_\alpha|\alpha\rangle_R\rightarrow\sum_\alpha \phi_\alpha|\alpha\rangle_R|D_\alpha\rangle_{dr}. \tag{8}$$

本算法需要 $O(\log MN)$次操作来访问数据,N 为训练数据样本数,如式(9)、(10)所示:

$$\frac{1}{\sqrt{MN}}|i\rangle|j\rangle|0\rangle\xrightarrow{\text{QRAM}}\frac{1}{\sqrt{MN}}|i\rangle|j\rangle|\boldsymbol{x}_i^{(j)}\rangle\xrightarrow{\text{逆 QRAM}}\frac{1}{\sqrt{MN}}|0\rangle|0\rangle|\boldsymbol{x}_i^{(j)}\rangle\rightarrow\frac{1}{\sqrt{MN}}|\boldsymbol{x}_i^{(j)}\rangle, \tag{9}$$

$$|0\rangle|0\rangle\xrightarrow{H^{\otimes m}}\frac{1}{\sqrt{N}}\sum_{i=0}^{2^n-1}|i\rangle|0\rangle\xrightarrow{\text{QRAM}}\frac{1}{\sqrt{N}}\sum_{i=0}^{2^n-1}|i\rangle|\boldsymbol{y}_i\rangle\xrightarrow{H}\frac{1}{\sqrt{N}}\sum_{i=0}^{2^n-1}|0\rangle|\boldsymbol{y}_i\rangle\rightarrow|\boldsymbol{y}_i\rangle. \tag{10}$$

步骤 2:量子计数

量子计数是该算法的核心步骤,目的在于计算数据库的 N 个数据中包括事件$(\tilde{\boldsymbol{x}}^{(j)},\boldsymbol{c}_k)$的个数. 通过测量得到我们所需要的概率[17],如图 2 所示,具体计算步骤如式(11)所示:

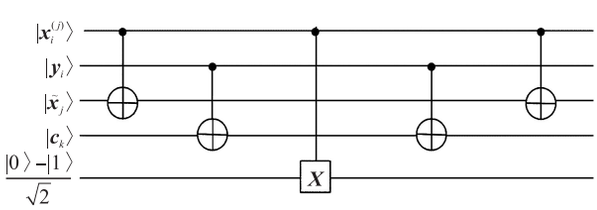

图 2　量子黑盒电路图

Fig. 2　The quantum oracle circuit diagram

$$|i\rangle|j\rangle|\boldsymbol{x}_i^{(j)}\rangle|\boldsymbol{y}_i\rangle|\tilde{\boldsymbol{x}}_j\rangle|\boldsymbol{c}_k\rangle\xrightarrow{\frac{|0\rangle-|1\rangle}{\sqrt{2}}}$$

$$|i\rangle|j\rangle|\boldsymbol{x}_i^{(j)}\rangle|\boldsymbol{y}_i\rangle|\tilde{\boldsymbol{x}}_j\oplus\boldsymbol{x}_i^{(j)}\rangle|\boldsymbol{c}_k\oplus\boldsymbol{y}_i\rangle\xrightarrow{\frac{|0\rangle-|1\rangle}{\sqrt{2}}}$$

$$\begin{cases}|i\rangle|j\rangle|\boldsymbol{x}_i^{(j)}\rangle|\boldsymbol{y}_i\rangle|\tilde{\boldsymbol{x}}_j\oplus\boldsymbol{x}_i^{(j)}\rangle|\boldsymbol{c}_k\oplus\boldsymbol{y}_i\rangle\dfrac{|1\rangle-|0\rangle}{\sqrt{2}}\\ |i\rangle|j\rangle|\boldsymbol{x}_i^{(j)}\rangle|\boldsymbol{y}_i\rangle|\tilde{\boldsymbol{x}}_j\oplus\boldsymbol{x}_i^{(j)}\rangle|\boldsymbol{c}_k\oplus\boldsymbol{y}_i\rangle\dfrac{|0\rangle-|1\rangle}{\sqrt{2}}\end{cases}. \tag{11}$$

此时,便实现了类似于 Grover 的黑盒操作.

$$O|i\rangle=\begin{cases}|i\rangle, & \boldsymbol{x}_i^{(j)}\neq\tilde{\boldsymbol{x}}^{(j)},\quad \boldsymbol{y}_i\neq\boldsymbol{c}_k,\\ -|i\rangle, & \boldsymbol{x}_i^{(j)}=\tilde{\boldsymbol{x}}^{(j)},\quad \boldsymbol{y}_i=\boldsymbol{c}_k.\end{cases}\tag{12}$$

构造与之对应的 G^{kj}算子,

$$G^{kj}=(2|\boldsymbol{\chi}_N\rangle\langle\boldsymbol{\chi}_N|-\boldsymbol{I}_N)O,\tag{13}$$

式中,$|\boldsymbol{\chi}_N\rangle=\dfrac{\sum_{i=0}^{N-1}|i\rangle}{\sqrt{N}}$,$\boldsymbol{I}_N$ 是 $N\times N$ 的单位矩阵.

G^{kj}算子拥有 2 个本征值 $\lambda_{\pm}=\mathrm{e}^{\pm 2\mathrm{i}\theta}(\mathrm{i}=\sqrt{-1})$及相对应的本征态 $|\psi_{\pm}\rangle=\dfrac{|\psi_0\rangle\mp\mathrm{i}|\psi_1\rangle}{\sqrt{2}}$.

那么,在 G^{kj}算子的本征态空间上重新描述$|\psi\rangle$,可得:

$$|\psi\rangle=\frac{\mathrm{e}^{\mathrm{i}\theta^{kj}}|\psi_+\rangle+\mathrm{e}^{-\mathrm{i}\theta^{kj}}|\psi_-\rangle}{\sqrt{2}}.\tag{14}$$

然后借助辅助粒子进行相位估计[3],可得:

$$|0\rangle|\psi_0\rangle\xrightarrow{\text{相位估计}}\frac{\sum_{j=1}^{M}\sum_{k=1}^{K}(\mathrm{e}^{\mathrm{i}\theta^{kj}}|\tilde{\theta}^{kj}\rangle|\psi_+\rangle+\mathrm{e}^{-\mathrm{i}\theta^{kj}}|\pi-\tilde{\theta}^{kj}\rangle|\psi_-\rangle)|\tilde{\boldsymbol{x}}_j\rangle|\boldsymbol{c}_k\rangle}{\sqrt{2MK}}.\tag{15}$$

步骤 3:投影测量

对第 2 量子寄存器和第 3 量子寄存器进行投影测量,当测量结果为$|\tilde{\boldsymbol{x}}_j\rangle$和$|\boldsymbol{c}_k\rangle$时,第 1 量子寄存器就塌缩到$|\tilde{\theta}^{kj}\rangle$. 然后,通过对该量子的测量,即可获得 $\tilde{\theta}^{kj}$的信息. 显然,获得该测量结果的概率为$\dfrac{1}{2MK}$. 由于,算法共有 M 个属性和 2 个类别,因此,进行 $O(2M)$次上述操作,就可获得所需信息. 为了进一步提高算法效率,可采用幅度放大技术[18],将操作次数减少到 $O(\sqrt{M})$.

步骤 4:计算类别

通过测量 $\tilde{\theta}^{kj}$得到 $P(\tilde{\boldsymbol{x}}^{(j)},\boldsymbol{c}_k)$,从而得到 $P(\boldsymbol{Y}=\boldsymbol{c}_k)\prod_{j=1}^{M}P(\boldsymbol{X}^j=\boldsymbol{x}^{(j)}|\boldsymbol{Y}=\boldsymbol{c}_k)$. 进行对比即可得到待分类数据所属类别.

3　复杂度分析

这一节中,对上述量子算法的时间复杂度进行简要分析. 各步骤的时间复杂度如表 1 所示.

步骤 1,利用 QRAM 访问经典数据,制备初态,此步骤运行时间复杂度为 $O(\log MN)$. 步骤 2,利用黑盒 oracle 实现目标态的振幅翻转,对 G^{kj}算子进行相位估计,由文献[3]可知,其时间复杂度为 $O(\log N)$. 步骤 3,通过投影测量,成功估计特定 $\tilde{\theta}^{kj}$的概率为 $1/(2MK)$. 结合幅度放大技术,只需重复进行 $O(\sqrt{2M})$次操作即可得到 $\tilde{\theta}^{kj}$,其时间复杂度为 $O(\sqrt{M})$. 步骤 4,在计算所得的数值中进行对比即可找出最大数值代表的类别,时间复杂度为 $O(1)$. 综上所述,该算法的时间复杂度为 $O[M(\log MN+\sqrt{M})]$.

相对经典贝叶斯二元分类算法的时间复杂度 $O(NM^2+M^2)$而言,显然,本文所提算法取得了量子指数级加速的效果. 与时间复杂度为 $O\left[\dfrac{\tilde{k}^2M^{2.5}(\log^2M+\log N)}{\varepsilon^3}\right]$的 Shao 算法[11]进行对比,本文所提算法在时间复杂度上也略优于 Shao 算法. 此外,Shao 算法利用厄米矩阵块编码技术来实现量子加速的目的,这就要求数据矩阵满足厄米条件,大大限制了该算法的适用范围. 本文所提算法可以应用于任意数据矩阵,具有更高的普适性. 综上,本文所提算法可实现量子指数级加速,且可应用的数据集更为广泛.

表 1　复杂度分析

Table 1　The complexity analysis

算法流程	时间复杂度
步骤 1	$O(\log MN)$
步骤 2	$O(\log N)$
步骤 3	$O(\sqrt{M})$
步骤 4	$O(1)$
总计	$O[M(\log MN+\sqrt{M})]$

4 结论

本文将量子力学理论融入机器学习中,提出了一种量子贝叶斯分类算法,运用量子计数方法高效地估计出所需数据,并设计了相应的量子电路图. 分析表明所提的量子贝叶斯算法的时间复杂度为 $O[M(\log MN+\sqrt{M})]$,当数据样本维数较低($M\leqslant N$)时,获得指数级加速. 该算法与仅可应用于厄米矩阵的 Shao 算法相比,可适用于更为广泛的数据集.

[参考文献]

[1] 杨双波,韦栋. 周期受击简谐振子系统的经典与量子动力学[J]. 南京师大学报(自然科学版),2011,34(4):49-54.

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

[3] GROVER L K. Quantum mechanics helps in searching for a needle in a haystack[J]. Physical review letters,1997,79(2):325-328.

[4] WIEBE N,KAPOOR A,SVORE K M. Quantum algorithms for nearest-neighbor methods for supervised and unsupervised learning[J]. Quantum information & computation,2014,15(3):318-358.

[5] 陈汉武,高越,张军. 量子 K-近邻算法[J]. 东南大学学报(自然科学版),2015,45(4):647-651.

[6] LU S,BRAUNSTEIN S L. Quantum decision tree classifier[J]. Quantum information processing,2014,13(3):757-770.

[7] REBENTROST P,MOHSENI M,LLOYD S. Quantum support vector machine for big data classification[J]. Physical review letters,2014,113(13):130503-130508.

[8] BISHWAS A K,MANI A,PALADE V. An all-pair quantum SVM approach for big data multiclass classification[J]. Quantum information processing,2018,17(10):1-16.

[9] YU C H,GAO F,WANG Q L,et al. Quantum algorithm for association rules mining[J]. Physical review A,2016,94(4):042311-042319.

[10] 吴嵘,张姣玲,刘小兰. 结合变异机制和量子 PSO 的关联规则挖掘算法[J]. 山东科技大学学报(自然科学版),2020,39(2):95-104.

[11] SHAO C P. Quantum speedup of bayes' classifiers[J]. Journal of physics A:mathematical and theoretical,2020,53(4):045301-045328.

[12] 邵晓根,鞠训光,胡局新,等. 基于改进权重的贝叶斯推理和 TFIDF 算法文本主题词提取研究[J]. 南京师大学报(自然科学版),2014,37(1):57-60,65.

[13] 汤胜道,殷世茂. 正态分布下参数的模糊贝叶斯估计(英文)[J]. 南京师大学报(自然科学版),2015,38(1):13-20.

[14] 张永军,刘金岭. 一种改进的高效贝叶斯短信文本分类器[J]. 南京师范大学学报(工程技术版),2014,14(3):70-74.

[15] 金文梁. 量子搜索算法的多相位关系研究[J]. 计算机学报,2012,35(7):1440-1447.

[16] GIOVANNETTI V,LLOYD S,MACCONE L. Quantum random access memory[J]. Physical review letters,2008,100(16):160501-160506.

[17] 张焕国,毛少武,吴万青,等. 量子计算复杂性理论综述[J]. 计算机学报,2016,39(12):2403-2428.

[18] BRASSARD G,HOYER P,MOSCA M,et al. Quantum amplitude amplification and estimation[J]. Contemporary mathematics,2002,305:53-74.

[责任编辑:丁 蓉]