

引用格式:赵超,牛伟纳,杨俊闯. 基于量子计算的分类和聚类算法综述[J]. 微电子学与计算机,2020,37(8):1-5. [ZHAO C, NIU W N, YANG J C. A survey on quantum classification and clustering algorithms [J]. Microelectronics & Computer,2020, 37(8):1-5.]

基于量子计算的分类和聚类算法综述

赵 超¹,牛伟纳²,杨俊闯¹

(1 河北工程大学 信息与电气工程学院,河北 邯郸 056038;

2 四川大学 网络空间安全学院,四川 成都 610207)

摘 要:越来越多的研究表明,借助量子计算技术可以提高有监督分类算法和无监督聚类算法的计算效率,甚至是学习精度.通常采用的方法有:基于量子理论将经典信息转换为量子态的形式存储起来,用量子态来表示所有样本;以量子态之间的距离替代样本数据之间的经典距离,形成新的相似度来度量样本数据间的相似性等.通过理论和模拟验证表明,量子计算可以实现对经典机器学习算法的加速.最后,总结了量子机器学习技术的优势和目前所存在的问题,并展望了未来该领域的发展趋势.

关键词:量子计算;机器学习;分类;聚类

中图分类号: TP301.6

文献标识码: A

文章编号: 1000—7180(2020)08—0001—05

A survey on quantum classification and clustering algorithms

ZHAO Chao¹, NIU Wei-na², YANG Jun-chuang¹

(1 School of Information & Electrical Engineering, Hebei University of Engineering, Handan 056038, China;

2 College of Cybersecurity, Sichuan University, Chengdu 610207, China)

Abstract: A lot of work shows that quantum computing can speed up supervised and unsupervised learning algorithms, even improving the performance of these algorithms. Usually the following methods are used: based on quantum theory, it encodes classical information into quantum state, and then represents all of samples in dataset by one or several quantum states; It quantifies the similarity of two samples by measuring the distance between the quantum states through a quantum algorithm. The theoretical and simulation results show that quantum computation can accelerate classical machine learning algorithms. Finally, it summarizes the advantages of quantum machine learning technology and the existing problems, and looks forward to the future development trend of this field.

Key words: quantum computing; machine learning; classification; clustering

1 引言

近年来,越来越多的研究表明量子计算技术可以对经典算法进行显著加速.在机器学习领域,同样可以得到类似的结论.因此,结合了量子计算和机器学习的量子机器学习领域得到了广泛的关注,涌现出了很多量子机器学习算法^[1-3].一方面,研究者们

希望通过量子计算解决机器学习的运算效率问题,降低传统机器学习算法的复杂度;另一方面,探索使用量子力学的性质,创造出可以真正运行在量子环境上的量子机器学习算法.量子计算是量子力学与计算机科学结合的产物.量子计算机可以展现出比传统计算机强大得多的性能.随着数据的海量增加,在当今大数据时代量子机器学习将发挥无与伦比的

收稿日期:2019—11—05;修回日期:2019—12—17

基金项目:国家自然科学基金项目(61902262);河北省高等学校科学技术研究项目(QN2018109);邯郸市科学技术研究与发展规划计划项目(19422031008-15)

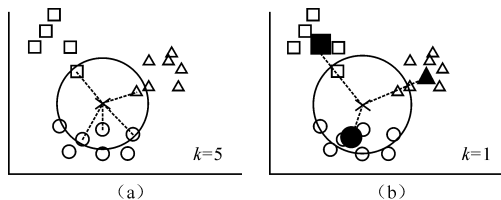
优势^[4-5].

传统机器学习主要可以分为三类:有监督学习、无监督学习和增强学习. 量子计算有助于研究有监督分类和无监督聚类问题, 本文将重点关注和总结该领域的进展. 例如将经典信息转换为量子态的形式存储起来, 用量子态来表示所有样本; 以量子态之间的距离替代样本数据之间的距离, 形成新的相似度量等.

2 量子有监督分类算法

2.1 量子 K 近邻算法

K 近邻(K-Nearest Neighbours, KNN)是一种传统有监督分类的算法, 它根据待分类数据与样本数据(也称为训练样本)之间的相似度进行分类. 相似度需要根据实际问题定义, 通常是通过距离来估量相似度. 它的主要思想是: 一个待分类样本, 在特征空间中如果与其最近邻的 K 个样本中大多数属于某一个类别, 则该待分类样本也属于这个类别, 如图 1(a)所示^[2]. 通常 K 值的选择并不容易, 而且很大程度上影响分类结果. 也可以计算出每个类别里的一个平均样本, 称为该类别的质心(centroid), 评估待分类样本与每个质心的距离, 距离最短的那个所属类别即为待分类样本的类别, 如图 1(b)所示^[2].



(a) 当 $k=5$ 时, 找到与待分类样本 $\times$ 最近的 5 个训练样本, 可知 $\times$ 所属于圆形类别; (b) 计算出每个类别的质心(阴影样本), 可知 $\times$ 与圆形阴影质心的距离最短, 即 $\times$ 所属于圆形类别. 此时可认为 $k=1$.

图 1 KNN 算法数据分类方式

在特征空间中, 样本即为一个个多维向量. 因此, 常用的相似度标准有: 比较两个样本的内积、欧氏距离、汉明距离、曼哈顿距离、夹角余弦距离等. 当想要把传统 KNN 算法转换为量子版本, 通常聚焦在如何使用量子计算来评估两个向量之间的距离, 以替代两个向量间的经典距离. 为此, 首先需要使用量子态来表示样本向量. Aïmeur 等人^[6]提出使用保真度(fidelity) $\langle a|b \rangle$ 作为两个量子态 $|a\rangle$ 和 $|b\rangle$ 的相似度. 保真度可以通过一个简单的 Swap 量子电路来得到, 如图 2 所示.

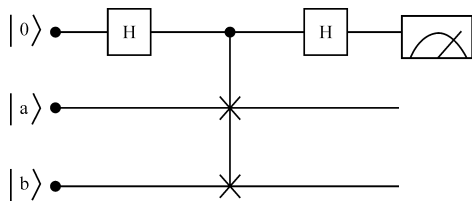


图 2 Swap 量子电路

其中 H 代表 Hadamard 门, 电路是否起作用由最上面的量子比特控制, 即电路最上面的量子线对应的是控制比特, 初始状态设置为 $|0\rangle$. 下面两个量子线对应的是受控量子比特. 第一个 H 门将控制比特 $|0\rangle$ 转换为一个叠加态 $(1/\sqrt{2})(|0\rangle + |1\rangle)$. 当控制比特的状态为 $|1\rangle$ 时, Swap 量子电路对于状态 $|a\rangle$ 和 $|b\rangle$ 进行交换操作. 经过第二个 H 门后可得到量子态

$$\begin{aligned} & \frac{1}{2} |0\rangle \langle a, b| + |b, a\rangle + \\ & \frac{1}{2} |1\rangle \langle a, b| - |b, a\rangle \end{aligned} \quad (1)$$

对最上面的控制比特进行测量, 得到状态 $|0\rangle$ 的概率为

$$p(|0\rangle) = \frac{1}{2} + \frac{1}{2} |\langle a|b \rangle|^2 \quad (2)$$

由上式可知, 当得到状态 $|0\rangle$ 的概率为 $1/2$ 时, 说明状态 $|a\rangle$ 和 $|b\rangle$ 无保真度(即 $|a\rangle$ 和 $|b\rangle$ 是正交的). 当得到 $|0\rangle$ 的概率为 1 时, 表明 $|a\rangle$ 和 $|b\rangle$ 有最大的保真度.

基于 Swap 量子电路, Lloyd 等人^[7]通过量子测量技术引出一个新的方法来度量两个 n 维实向量 $\vec{a}$ 和 $\vec{b}$ 的距离. 具体来说, 计算两个状态 $|\psi\rangle = (1/\sqrt{2})(|0, a\rangle + |1, b\rangle)$ 和 $|\phi\rangle = (1/\sqrt{2})(|a|0\rangle - |b|1\rangle)$ (其中 $Z = |\vec{a}|^2 + |\vec{b}|^2$) 的内积. 即借助 Swap 量子电路计算 $|\langle \psi | \phi \rangle|^2$. 其中, 状态 $|\psi\rangle$ 和 $|\phi\rangle$ 可以被有效的制备. 该方法的关键思想是, 将一个向量 $\vec{x}$ 的经典长度编码到量子态自身的内积上 $\langle x | x \rangle = |\vec{x}|^{-1} \vec{x}^2$ ^[7,8]. 根据此定义, 就可以得到 $|\vec{a} - \vec{b}|^2 = Z |\langle \phi | \psi \rangle|^2$. 至此, 两个向量 $\vec{a}$ 和 $\vec{b}$ 的距离可由两个制备态 $|\psi\rangle$ 和 $|\phi\rangle$ 的 Swap 量子电路得到, 从而使用了量子计算得到了两个向量间的经典距离. 基于此距离计算方法, Lloyd 等人提出了相应的量子最近邻质心(nearest-centroid)算法: 计算待分类样本与每个质心的距离, 进而决定待分

类样本的类别。Lloyd 等人认为即使考虑到要制备相应的量子态,该量子算法也比对应的经典算法更加高效。

同样是基于 Swap 量子电路来计算两个向量的内积 Wiebe 等人^[9]提出另一种距离测量方式。借助于量子态,他们使用了不同的表示经典信息的方式。

假定两个 n 维向量 $\vec{a}$ 和 $\vec{b}$, 其特征元素(向量元素)分别为 $a_j = |a_j|e^{i\alpha_j}$, $b_j = |b_j|e^{i\beta_j}$, $j = 1, 2, \dots, n$. 并且令 $r_{\max}$ 为训练样本集中的特征元素上界, d 为非零特征的数量上界. 该方法的核心思想是, 将经典信息中的非零数据编码到量子态的概率幅值上,

$$|A\rangle = \frac{1}{\sqrt{d}} \sum_j |j\rangle \left(\sqrt{1 - \frac{|a_j|^2}{r_{\max}^2}} e^{-i\alpha_j} |0\rangle + \frac{a_j}{r_{\max}} |1\rangle \right) \quad (3)$$

$$|B\rangle = \frac{1}{\sqrt{d}} \sum_j |j\rangle |1\rangle \left(\sqrt{1 - \frac{|b_j|^2}{r_{\max}^2}} e^{-i\beta_j} |0\rangle + \frac{b_j}{r_{\max}} |1\rangle \right) \quad (4)$$

对状态 $|A\rangle$ 和 $|B\rangle$ 进行 Swap 量子操作后, 根据等式(2), 控制比特为状态 $|0\rangle$ 的概率为

$$P(|0\rangle) = \frac{1}{2} + \frac{1}{2} \left| \frac{1}{dr_{\max}^2} \sum_i a_i b_i \right|^2 \quad (5)$$

则向量 $\vec{a}$ 和 $\vec{b}$ 的内积为

$$\left| \sum_i a_i b_i \right|^2 = d^2 r_{\max}^4 (2P(|0\rangle) - 1).$$

此外, Wiebe 等人^[9]基于欧式距离(Euclidian distance)提出了一个最近邻质心算法, 但实验结果表明基于欧式距离的方法并不优于内积方法。

2.2 其它量子分类算法

研究量子 KNN 算法有助于我们理解量子计算如何与机器学习领域有效结合。除此之外, 在有监督量子分类算法里, 量子版本的支持向量机(Support Vector Machine, SVM)也得到了越来越多的研究。传统 SVM 是通过寻找最大间隔分类超平面进行分类。2014 年, Rebentrost 等人^[8]提出了量子 SVM 算法。他们认为内积的运算在量子计算机上可以更快的执行, 从而借助量子算法解决训练样本内积运算问题。在物理实现方面, 量子 SVM 近年来也取得了进展^[10], 这也表明了量子计算与机器学习理论结合的可行性。

人工神经网络(Artificial Neural Network, ANN)是一种通过模拟生物神经网络的计算模型, 应用十分广泛。一个经典的 ANN 是基于多层次的

网状结构, 如图 3 所示^[1]。

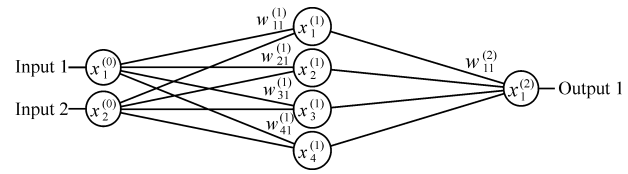


图 3 一个人工神经网络示意图

图中的节点称为神经元(neurons), 节点之间通过带有权重的突触(synapses)来连接。每个神经元都有一个激活函数, 该函数通过上层神经元的值以及相应权重, 从而决定了当前神经元的值。图 3 的第一层为输入层, 最后一层为输出层, 中间层(可以多层)为隐含层。

ANN 将训练样本输入网络, 根据网络处理后的输出结果, 优化以网络权重为参数的损失函数。经过训练后的神经网络, 输出结果与分类标签的误差最小。有研究指出生物神经网络与量子效应有着潜在的关系, 故产生了很多量子神经网络的研究: 试图探索整合了神经网络动力学性特征的量子线路, 进而利用神经网络进行量子计算^[11]。最早在 1996 年, 就有研究者提出了量子 ANN。随后二十多年, 量子神经网络的研究大量涌现, 并取得了显著的进展。2014 年, Schuld 等人^[12]对现有的量子神经网络研究做了比较系统的分析和总结: 文中指出量子神经网络应当遵循一些必备条件, 但是现有的绝大多数方法都无法同时满足。

此外还有很多量子分类算法, 例如量子决策树(Quantum Decision Tree)算法、利用贝叶斯网络进行量子态分类、以及由量子神经网络衍生出的量子行走(Quantum walks)和量子深度学习等等, 具体可以参考文献[1-3]。

3 量子无监督聚类算法

3.1 量子 K-means 算法

聚类是指将一些没有提前分好类别的样本数据, 根据学习算法自行找到数据间的潜在规则和相似度, 从而将样本数据划分为若干类。因此不同与分类算法, 聚类算法不需要带有分类标签的训练样本集, 算法应该知道如何计算样本数据之间的相似度并且将它们自动归类。而相似度的评估通常也是基于样本数据之间的距离。在传统无监督聚类算法中, K-means 算法的使用最为普遍: 每次聚类时找出每个类别的平均值样本即质心, 计算待聚类样本与每个质心的距离, 其中距离最短的质心所属类别即为

待聚类样本的类别. 然后更新所有的质心位置, 最后算法收敛于所有质心位置不变. 当然, 在最初的聚类开始前, 需要初始随机选取 K 个质心. 以上过程通常也被称为最近邻质心算法, 如图 4 所示^[2].

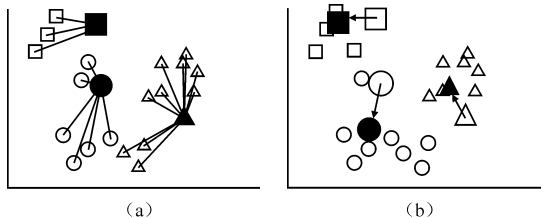


图 4 K-means 算法的聚类过程

由于初始质心是随机选取, 因此算法可能会导致局部最优. 另外, 在不了解样本数据的情况下如何选取参数 K 值, 这些都是 K-means 算法所面临的问题. Lloyd 等人^[7]构造了量子版的最近邻质心聚类算法. 该算法核心思想仍然是比较量子态之间的距离. 首先将输入样本向量 u 转化为量子态 $|u\rangle = u/|u|$, 进而实向量的距离比较再次转化为量子态间的距离比较. 该方法表明, 一个待聚类样本 $|u\rangle$ 与包含 n 个样本 $\{|v_j^c\rangle\}$ 的聚类 C 的距离, 可在量子计算机上高效的计算. 为此构造状态

$$|\psi\rangle = \frac{1}{\sqrt{2}} \left(|u\rangle |0\rangle + \frac{1}{\sqrt{n}} \sum_{j=1}^n |v_j^c\rangle |j\rangle \right) \quad (6)$$

且对如下状态进行 Swap 量子操作

$$|\phi\rangle = \frac{1}{\sqrt{Z}} \left(|u\rangle |0\rangle - \frac{1}{\sqrt{n}} \sum_{j=1}^n |v_j^c\rangle |j\rangle \right) \quad (7)$$

式中, $|u|$ 为实向量 u 的范数; $Z = |u|^2 + (1/n) \sum_j |v_j^c|^2$ 为归一化系数. 同理等式 (1) 和 (2), 待聚类样本 $|u\rangle$ 与聚类 C 中平均样本 (质心) 的距离, 可由成功地进行了 Swap 量子操作的概率得到. 上述算法步骤对于样本空间中每个聚类都执行一次, 找到一个与 $|u\rangle$ 最短距离的类, 进而将待聚类样本 $|u\rangle$ 划分为该聚类内部.

K-means 算法中初始质心的选取十分重要, 可能会导致局部最优. 通常原则是将初始质心位置尽可能地稀疏分布于整个样本空间中, 即各初始质心位置之间距离尽可能大. 为此 Lloyd 等人^[7]还提出一个量子绝热计算的方法来解决寻找初始质心的优化问题. 量子绝热计算同样是基于量子 (酉变换) 操作来进行状态间的演化, 研究表明该方法可以应用到量子机器学习领域^[13].

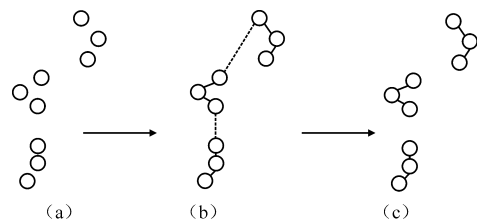
物理实验方面, 量子 K-means 算法也取得了重要突破. 2014 年, 中科大潘建伟院士团队借助光子量子计算机, 通过物理实验验证了 Lloyd 提出的量子 K-means 算法^[14]. 该实验表明在量子计算机上实现诸如向量间的内积和距离运算, 其效率优于传统算法. 这也预示着量子机器学习算法在处理大数据时有着明显的优势.

3.2 其它量子聚类算法

在传统无监督机器学习领域, K-median 算法、分裂聚类、以及最小生成树算法也都有着广泛的应用. 其中 K-median 算法类似于 K-means, 只是在选取聚类中心时的方法有所不同. Aïmeur 等人^[15]提出了量子 K-median 算法. 该算法基于两步量子计算子程序: 首先, 借助黑箱算子 (Oracle) 计算两个量子态间的距离, 进而可得到每个量子态与一个聚类中所有状态的距离总和; 基于最小值搜索算法, 找到上述距离总和的一个最小值, 从而将对应的量子态选为该聚类新的聚类中心 (称为 median).

K-median 算法过量子搜索算法进行加速, 但本质上属于分裂聚类算法, 对传统分裂聚类算法的主体框架并未进行改变. 分裂聚类作为一类简单常见的无监督学习算法, 首先将所有样本数据视为一个聚类, 然后不断分裂, 直至每个聚类不能再分裂. 若处理大量样本且样本数据维度较高时, 经典分裂聚类算法的效率就会较低. 因此以量子 K-median 为代表的量子分裂算法, 通过结合量子计算步骤进行了明显加速, 也为其余的量子学习算法提供了较好的借鉴意义.

最小生成树算法的目标是, 在特征空间中找到一个连接所有样本数据点 (且每个数据点仅被连接一次) 的最短路径, 称为最小生成树, 如图 5 所示^[1].



(a) 特征空间中的原始样本数据点; (b) 生成数据点的最小生成树; (c) 通过消除最小生成树中 $k-1$ 条最长的边 (图 5 (b) 中虚线), 可生成 k 个子类.

图 5 最小生成树算法聚类过程

一旦找到了最小生成树, 消除其中 $k-1$ 条最长的连接边, 即可将样本数据点划分为 k 个子类, 如图

5(c)所示(此时 $k=3$)。依靠经典计算机找到一个最小生成树路径,通常需要 $O(N^2)$ 时间, N 为特征空间中样本点的个数。但是借助 Oracle 算子和 Grover 搜索算法的量子生成树算法,可对传统算法进行加速。该方法还可用于在一些图论问题中找到最近邻的数据点。

有另一类不同于上文中的量子机器学习算法,特点是探寻量子力学效应和传统机器学习的相似点,将物理过程应用于传统问题的求解,产生出全新的机器学习算法,该类算法全部过程都可在经典计算机上实现。基于量子力学的聚类算法就是其中的代表性研究,在此不再详述。

4 结束语

量子机器学习是结合了量子计算和传统机器学习的交叉研究领域,本文聚焦于其中的有监督分类和无监督聚类问题。量子机器学习的研究方法,大体上可分为两类。第一类是探索相应的量子算法来替代经典机器学习算法,通过计算效率、算法复杂度和学习精度来体现量子算法的优越性。本文中量子 K 近邻和量子 K-means 为代表的大部分算法,都属于此类方法。它们借助量子计算步骤来加速传统的、低效的距离计算。另一类方法是用量子理论的概率思想来描述随机过程,例如量子神经网络和量子决策树等。

由于量子计算本身的高并行性等特点,尤其是对大数据的处理、分析和挖掘能力,使得量子机器学习研究给人以无限的遐想。但是该新兴领域还处于初步发展阶段,需要理性看待。首先该领域目前还没有一个完备的理论框架。再者通用的量子计算机设备还在研发过程中,虽然众多量子算法能从理论和模拟上证明其加速效果,但大部分还没有在量子环境下进行实验验证。量子机器学习这一领域的研究,毫无疑问还面临着诸多问题和挑战,但并不妨碍研究者们对它的前景充满信心。

参考文献:

- [1] ADCOCK J, ALLEN E, DAY M, et al. Advances in quantum machine learning[EB/OL]. (2015) Preprint-at <https://arxiv.org/abs/1512.02900>.
- [2] SCHULD M, SINAYSKIY I, PETRUCCIONE F. An introduction to quantum machine learning[J]. Contemporary Physics, 2014, 56(2): 172-185.
- [3] HUANG Y M, LEI H, LI X Y. A survey on quantum machine learning[J]. Chinese Journal of Computers, 2018, 41(1): 145-163.

- [4] WANG S H, LONG G L. Big data and quantum computation[J]. Chinese Science Bulletin, 2015 (5): 499-508.
- [5] GAO Q, ZHANG F L, WANG R J, et al. Trajectory Big Data: A Review of Key Technologies in Data Processing[J]. Journal of Software, 2017, 28(4): 959-992.
- [6] AiMEUR E, BRASSARD G, GAMBS S. Machine learning in a quantum world[J]. Lecture Notes in Computer Science, 2006, 4013: 431-442.
- [7] LLOYD S, MOHSENI M, REBENTROST P. Quantum algorithms for supervised and unsupervised machine learning. [EB/OL]. (2013) Preprint at <https://arxiv.org/abs/1307.0411v2>.
- [8] REBENTROST P, MOHSENI M, LLOYD S. Quantum support vector machine for big data classification[J]. Physical review letters, 2014, 113(13): 130503.
- [9] WIEBE N, KAPOOR A, SVORE K. Quantum algorithms for nearest-neighbor methods for supervised and unsupervised learning[J]. Quantum Information & Computation, 2014, 15(3): 318-358.
- [10] LI Z, LIU X, XU N, et al. Experimental realization of a quantum support vector machine[J]. Physical Review Letters, 2015, 114(14): 140504.
- [11] SILVA A S, OLIVEIRA W R, LUDERMIR T B. Classical and superposed learning for quantum weightless neural networks[J]. Neurocomputing, 2012, 75(1): 52-60.
- [12] SCHULD M, SINAYSKIY I, PETRUCCIONE F. The quest for a quantum neural network[J]. Quantum Information Processing, 2014, 13(11): 2567-2586.
- [13] PUDENZ K L, LIDAR D A. Quantum adiabatic machine learning[J]. Quantum Information Processing, 2013, 12(5): 2027-2070.
- [14] CAI X D, WU D, SU Z E, et al. Entanglement-based machine learning on a quantum computer[J]. Physical Review Letters, 2014, 114(11): 110504.
- [15] AiMEUR E, BRASSARD G, GAMBS S. Quantum clustering algorithms[C]// Proc of the 24th international conference on machine learning, New York: [s. n.], 2007: 1-8.

作者简介:

赵超男, (1986-), 博士, 讲师。研究方向为量子计算、机器学习。E-mail: zc57@163.com

牛伟纳女, (1990-), 博士。研究方向为网络安全、流量分析。

杨俊闯男, (1995-), 硕士研究生。研究方向为机器学习。