

基于一维三态量子游走的量子聚类算法

徐永振 郭躬德 蔡彬彬 林 崧

(福建师范大学数学与计算机科学学院 福州 350007)

(福建师范大学网络安全与密码技术福建省高校重点实验室 福州 350007)

摘 要 量子游走具有与经典随机游走不同的特性,因此它已经被用来解决包括元素区分、组合优化、图同构等问题。考虑量子游走和聚类两个领域,提出了一个基于一维三态离散量子游走的聚类算法。在该算法中,将数据点看作游走粒子;然后,这些粒子执行三态量子游走,接着根据粒子的测量结果更新数据点的属性值;最后,属于同一簇的数据点将会聚集,而属于不同簇的数据点将会分离。仿真实验结果表明了所提算法的有效性。

关键词 量子游走,无监督学习,数据聚类,量子计算

中图分类号 TP18 文献标识码 A DOI 10.11896/j.issn.1002-137X.2016.3.016

Quantum Clustering Algorithm Based on One-dimensional Three-state Quantum Walk

XU Yong-zhen GUO Gong-de CAI Bin-bin LIN Song

(School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, China)

(Key Laboratory of Network Security and Cryptography, Fujian Normal University, Fuzhou 350007, China)

Abstract As compared to classical random walk, quantum walk exhibits some remarkably different features. Thus, it has been exploited to solve some kinds of problems, e. g. element distinctness, combinatorial optimization, graph matching problem and so on. Considering the two related research fields of quantum walks and data cluster analysis, a quantum clustering algorithm based on the one-dimensional three-state quantum walk was proposed. In this algorithm, each data point is regarded as one particle. Then, these particles perform the three-state quantum walk. After that, according to the measurement results, the attribute values of the corresponding data points are updated. At last, the data points belonging to the same group will gather together, while those belonging to different group will be separated. The simulation results demonstrate that the proposed algorithm is effective.

Keywords Quantum walk, Unsupervised learning, Data clustering, Quantum computation

1 引言

量子计算是量子力学与计算机科学相结合的一门交叉学科,它利用量子系统的一些独特性质来解决计算机科学领域中的若干问题,对经典计算实现二次加速。例如,1984年数学家 Shor 提出了著名的 Shor 大数质因子分解算法^[1],该算法可在多项式时间内解决这一 NP 难问题。此外,人们还设计了许多高效的量子算法,如 Grover 量子搜索算法^[2]、量子绝热算法^[3]等。数据挖掘为知识发现提供手段,可以从包含大量的、多维的、不全的、含噪声的数据的集合中抽取隐含的、先前未知的、对决策有潜在价值的规则^[4]。随着量子技术的不断发展,研究者们开始将量子计算应用到数据挖掘领域,并取得了一些成果^[5,6]。

利用量子游走来设计量子算法是近年来的一个研究热点^[7]。量子游走是经典随机游走在量子领域的延伸。与经典随机游走相似,量子游走中粒子状态的瞬时状态是由所处位

置的概率分布来描述的,在量子游走中粒子是处在很多位置的某一叠加态中。量子游走在量子搜索算法中的显著优点推动了它的深入研究,产生了许多有趣的性质。此外,基于量子游走的量子通用计算同样突出了它的巨大潜力^[8]。因此,人们希望利用量子游走这一新型的量子计算工具来实现数据挖掘方面的一些突破。2011年, Li 等人^[9]首次把量子游走引入聚类算法中,把多维情形分解到一维情形,然后应用一维二态离散量子游走的测量结果来移动数据点,数据集上相似的数据点最终会自动聚集到一起。

受到该算法思想^[9]的启发,本文提出一个基于一维三态离散量子游走的聚类算法。在该算法中,把数据点看作游走粒子,首先构造粒子游走空间,然后对粒子执行若干步的三态量子游走,最后根据对量子态的测量结果选择对应的更新规则来移动数据点。随着不断的移动,数据点在空间中自动地聚集。

本文第2节介绍本文所涉及的量子游走的一些基础知

到稿日期:2015-03-25 返修日期:2015-06-08 本文受国家自然科学基金项目(61202451),福建省教育厅重点项目(JA12062)资助。

徐永振(1988—),男,硕士生,主要研究方向为量子信息与量子计算、量子人工智能等;郭躬德(1965—),男,博士,教授,博士生导师,主要研究方向为数据挖掘、机器学习等;蔡彬彬(1992—),男,硕士生,主要研究方向为量子密码等;林 崧(1979—),男,博士,副教授,主要研究方向为量子信息与量子计算、信息安全等,E-mail: lins95@fjnu.edu.cn(通信作者)。

识;第3节对本文所提出的一个基于一维三态离散量子游走的量子聚类算法进行详细描述;第4节给出实验仿真结果和分析;最后总结全文并给出下一步的研究方向。

2 量子游走

2.1 基本概念

比特是经典信息和经典计算机中的最小单位,通常用0和1表示。量子计算和量子信息在经典比特的概念上推出了量子比特。量子比特由量子态 $|0\rangle$ 和 $|1\rangle$ 来表示,与经典比特不同的是量子比特可以是叠加态:

$$|\psi\rangle = a|0\rangle + b|1\rangle \quad (1)$$

其中, a 和 b 分别表示出现 $|0\rangle$ 和 $|1\rangle$ 的概率幅,满足 $a^2 + b^2 = 1$ 。

如果有多个量子比特,可以把量子态用直积的形式写在一起。假设有两个量子比特,第一个量子比特为 $|0\rangle$,第二个量子比特为 $|1\rangle$,则由两个量子比特组成的系统的量子态可以表示为 $|0\rangle \otimes |1\rangle$ 。

一个独立的量子系统可以用一个量子态来表示,而这个量子态可以看成 Hilbert 空间中的复矢量。如果一个量子系统是封闭的,那么这个系统的演化必定是幺正的,也就是说一个系统的两个不同时间的量子态可以用一个幺正算符联系起来。量子计算就是对一个量子系统进行一系列幺正操作后,根据测量的结果来解决问题。

2.2 一维三态量子游走

量子游走主要分为离散和连续两种。本文中研究的模型是带硬币翻转的离散量子游走,它所对应的 Hilbert 空间可以表示为 $H = H_c \otimes H_p$ 。其中 H_p 表示游走粒子的位置空间,是无限但维度可数的希尔伯特空间。通常 H_p 的计算基的向量形式表示粒子的位置。 H_c 是硬币翻转算子作用的空间,维度 d 表示游走粒子能选择的方向数目。整个量子系统的幺正演化操作 U 由独立的两部分组成,即翻转硬币操作和有条件的置换操作: $U = S \cdot (C \otimes I)$ 。

量子游走的第一步是在硬币空间执行一个旋转操作 C ,这相当于经典随机游走中的抛硬币,通过这个操作得到一个硬币的叠加态。然后,移位操作 S 使粒子沿着由硬币决定的一条边移动到下一个邻近的点。带有硬币翻转的量子有许多变形,如一维直线游走、平面游走、超立方体游走等。本文所用的模型是一维直线游走中的三态量子游走。

对于一维三态量子游走^[10],其位置空间是 $H_p = \{|0\rangle, |\pm 1\rangle, \dots\}$,硬币空间是 $H_c = \{|0\rangle, |1\rangle, |2\rangle\}$ 。游走粒子根据3种硬币态来移动。移位操作 S 作用到基矢上可表示为:

$$S|0, x\rangle = |0, x-1\rangle, S|1, x\rangle = |1, x\rangle, S|2, x\rangle = |2, x+1\rangle \quad (2)$$

硬币翻转算符 C 的形式如下:

$$C = -\frac{1+s}{2}|0\rangle\langle 0| + \frac{n}{\sqrt{2}}|1\rangle\langle 0| + \frac{1-n}{2}|2\rangle\langle 0| + \frac{n}{\sqrt{2}}|0\rangle\langle 1| + s|1\rangle\langle 1| + \frac{n}{\sqrt{2}}|2\rangle\langle 1| + \frac{1-s}{2}|0\rangle\langle 2| + \frac{n}{\sqrt{2}}|1\rangle\langle 2| - \frac{1+s}{2}|2\rangle\langle 2| \quad (3)$$

其中, s 和 n 满足 $s^2 + n^2 = 1$ 。对初始状态为 $|0\rangle_p \otimes |1\rangle_c$ 的粒子执行一步三态量子游走,过程如下:

$$\begin{aligned} |0\rangle_p \otimes |1\rangle_c &\xrightarrow{C} \frac{n}{\sqrt{2}}|0\rangle_p \otimes |0\rangle_c + s|0\rangle_p \otimes |1\rangle_c + \frac{n}{\sqrt{2}}|0\rangle_p \otimes |2\rangle_c \\ &\xrightarrow{S} \frac{n}{\sqrt{2}}|-1\rangle_p \otimes |0\rangle_c + s|0\rangle_p \otimes |1\rangle_c + \frac{n}{\sqrt{2}}|1\rangle_p \otimes |2\rangle_c \end{aligned}$$

在游走过程中不进行测量,保持态的相干叠加性,经过若干步以后就会得到与经典随机游走不同的结果。图1是一维二态量子游走的几率分布,不同于经典的高斯分布,它有一种向左的趋势,这是由量子相干效应造成的。图2是一维三态量子游走的几率分布。它与一维二态量子游走的几率分布的一个最大不同是,经过若干步以后,发现粒子有很大的概率在初始位置,即所谓的“局部化”现象。本文所提出的算法正是利用这个特性,看作数据点的粒子在它的游走空间执行若干步三态量子游走,如果它的驻留概率很大,也就是说它和游走空间中的其它粒子很相似,那么粒子将有很大概率不移动或有很小的移动;如果它的驻留概率很小,也就是说它和游走空间中的其它粒子极不相似,那么粒子将有很大的概率跳跃。

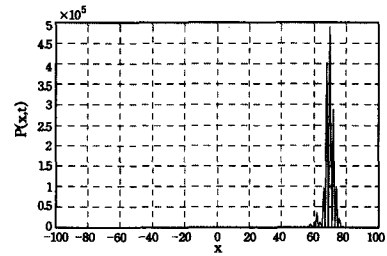


图1 一维两态量子游走的概率分布,初始态是 $|0, R\rangle$

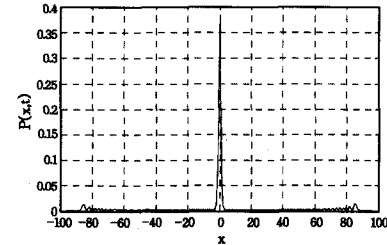


图2 一维三态量子游走的概率分布,初始态是 $|0\rangle \otimes \frac{1}{\sqrt{3}}(|0\rangle + |1\rangle + |2\rangle)$, $s = \frac{1}{\sqrt{2}}$

3 算法

首先对本文所提的算法进行简要的描述。假设有一个具有 m 个样本点的 d 维数据集 $DB = \{X_i = (X_i^1, X_i^2, \dots, X_i^d, \dots, X_i^d) | i=1, 2, \dots, m\}$,其中 X_i^k 为第 i 个样本点的第 k 维属性值。在本文的算法中,将数据集集中的每一个数据点看作有限 Hilbert 空间上的一个粒子,这样该粒子就可在空间中进行三态量子游走。在 t 步后,对该粒子进行投影测量,并根据测量结果按照一定的更新规则来对样本点的数据进行更改,进而实现样本点的移动。随着数据集集中的每个数据点随机地移动,最后相似的数据点会自动聚集在一起。

在设计算法的过程中存在以下几个问题需要解决:对于任意数据点如何构造它的游走空间,三态量子游走过程中的逗留概率 s 的设计,游走的步数 t 和数据点更新规则的构造。接下来就对这3个主要问题逐一进行描述。

3.1 游走空间

定义 1 给定一个 r , 对于任一数据点, 以自身为中心, 以 $c_1 \cdot r$ 为半径的范围内的数据点构成它的邻居集合 Γ_i , 以 $c_2 \cdot r$ 为半径的范围内的数据点构成它的近邻居集合 Φ_i , 其中 $c_1 > c_2$ 。

数据点的游走空间 H 中硬币空间和式(2)一样, 而位置空间和邻居集集合相关, 定义如下:

$$H_p = \{0, \pm 1, \pm 2, \dots, \pm |\Gamma_i|\}$$

其中, $|\Gamma_i|$ 是集合势。

3.2 逗留概率

三态量子游走的逗留概率 s 应满足: 粒子与它邻居集合中的粒子越相似, s 越趋近于 1。我们认为 s 与邻居的距离、数目等相关, 采用如下的定义形式:

$$s(i) = \frac{|\Phi_i| \sum_{k \in \Phi_i} \text{Deg}(k) \sum_{j \in \Phi_i} d(k, j)}{\omega |\Gamma_i| \sum_{j \in \Gamma_i} \text{Deg}(j) \sum_{j \in \Gamma_i} d(i, j)} \quad (4)$$

其中, ω 是参数, 默认值为 1, $|\Gamma_i|$ 和 $|\Phi_i|$ 是集合势, $\text{Deg}(\cdot)$ 是数据点的度, $d(\cdot, \cdot)$ 表示两个数据点的距离。

将 d 维空间上的量子游走分解为 d 个一维三态量子游走。对于每一维, 数据点有 3 个选择: 向左或向右一步, 或不发生改变。利用逗留概率 s 可以构造游走所用的硬币矩阵 C :

$$C = \begin{bmatrix} -\frac{1+s}{2} & \frac{n}{\sqrt{2}} & \frac{1-s}{2} \\ \frac{n}{\sqrt{2}} & s & \frac{n}{\sqrt{2}} \\ \frac{1-s}{2} & \frac{n}{\sqrt{2}} & -\frac{1+s}{2} \end{bmatrix} \quad (5)$$

其中, $s^2 + n^2 = 1$ 。

3.3 数据点更新

粒子经过 t 步量子游走后, 如果 s 足够大, 那么粒子落在原点附近的概率很大。 s 足够大也说明数据点和它的邻居内的点很相似, 数据点应该在邻域中移动。如果 s 足够小, 则说明数据点和周围点的平均相似度不大, 说明数据点应该向外大步移动。

定义 2 对于数据点每一维度 X_i^k , 以它为中心, 把它的邻居分为两部分, 大于它的为第一部分 $\max_i$, 小于它的为第二部分 $\min_i$ 。

第一部分的更新步长 $\Omega(\max_i) = \frac{X_j^k - X_i^k}{|\Gamma_i|}$, $j \in \Gamma_i$, 其中 X_j^k 是第一部分离 X_i^k 最远的点。

第二部分的更新步长 $\Omega(\min_i) = \frac{X_i^k - X_j^k}{|\Gamma_i|}$, $j \in \Gamma_i$, 其中 X_j^k 是第二部分离 X_i^k 最远的点。因此, 根据量子游走的位置测量结果 g , 定义如下的数据点更新规则:

$$X_i^k(t+1) = \begin{cases} X_i^k(t) + g \cdot \Omega(\max_i), & g > 0 \\ X_i^k(t), & g = 0 \\ X_i^k(t) + g \cdot \Omega(\min_i), & g < 0 \end{cases} \quad (6)$$

基于以上 3 个问题的说明, 下面对所提量子聚类算法的整个过程进行完整的描述。

算法 基于一维三态量子游走的量子聚类方法

1. 初始化参数设置和数据预处理;
2. 对于每个数据点;
3. 对应粒子执行一维三态量子游走;
4. 测量粒子;
5. 根据粒子测量结果, 对应数据点由式(6)进行更新;
6. 如果总更新步长小于给定阈值 c_3 , 停止算法;
7. 否则回到步骤 2 继续。

4 实验

为了说明算法的合理性和有效性, 在 4 个 UCI 数据集上进行实验评估, 并将其与 Li 等人的 1D-SCMS 算法^[9] 和 K-means^[12] 算法比较。首先简要介绍实验所用的 4 个数据集, 然后在聚类精度方面与其它算法进行比较, 最后给出结果分析。

4.1 实验数据

实验采用的 4 个 UCI 数据集的情况如表 1 所列。

表 1 实验数据的描述

数据集名称	实例个数	特征数目	标准类数
Iris	150	4	3
Glass	214	9	6
Ionosphere	351	32	2
Soybean	47	35	4

在数据预处理阶段, 由于所提算法实质是在每一个属性上进行运算的, 因此把数据的属性值映射到 $[0, 1]$ 区间来消除数值差异。本实验采用如下的距离计算公式:

$$d(X_i, X_j) = \exp\left(-\frac{\|X_i - X_j\|^2}{2\sigma^2}\right) \quad (7)$$

其中, X_i 和 X_j 是数据集中的实验数据, 一般取 $\sigma = 1$ 。

4.2 实验结果与分析

本文采用聚类精度来说明算法的有效性。聚类精度采用 Erkan 等人^[11] 提出的计算方法, 计算公式为:

$$r = \frac{\sum_{i=1}^N \lambda(\text{map}(\text{cluster}_i), c_i)}{N} \quad (8)$$

$$\lambda(\text{map}(\text{cluster}_i), c_i) = \begin{cases} 1, & \text{如果 } \text{map}(\text{cluster}_i) = c_i \\ 0, & \text{否则} \end{cases}$$

其中, cluster_i 是数据点 X_i 通过算法获得的类别号, c_i 是真实类别号。

最后给出 3 种算法在 4 个数据集上实验的聚类精度, 如表 2 所列, 本文的算法用 1D-3TQC 表示。从表 2 可以看出本文的算法在 Iris 等数据集取得了较高的聚类精度。这些结果表明所提出的算法是合理且有效的。

表 2 聚类精度的比较

数据集名称	1D-3TQC	K-means	1D-SCMS
Iris	95.4	89.3	94.7
Glass	62.3	47.2	63.1
Ionosphere	76.9	71.2	75.5
Soybean	92.6	47.2	91.5

4.3 参数 r 对算法的影响

算法中的参数有 4 个, 即 r, c_1, c_2, c_3 。下面以数据集 Iris 为例阐述这些参数对聚类效果的影响。对于 r, c_1, c_2 这 3 个参数来说, 它们的选择控制着粒子的一维三态量子游走, 进而

会影响聚类精度。这 3 个参数实际上主要是 r 有着决定作用。固定 c_1, c_2 的值, 研究聚类精度随着 r 的变化会有怎样的变化。在 $c_1=2, c_4=1, c_3=0.001$ 情况下, 实验结果如图 3 所示。从图 3 可以看出聚类精度并不是随 r 的增大而变大, 这是由于 r 很大时, 比如极端情形下, 每个数据点的邻居都包括数据集中其余的数据点, 在此情况下, 数据点的逗留概率很大, 则数据点的更新幅度很大, 进而会产生聚类错误。然而在 r 很小时, 数据点又几乎不进行移动。所以选择合适的 r 值很重要, 我们将在下一步工作中对此进行深入研究。

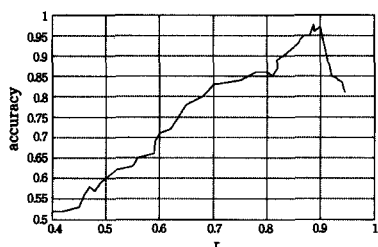


图 3 不同 r 的 Iris 的聚类精度

结束语 基于一维三态量子游走的“局部化”的特性, 把数据点看作游走空间中的粒子, 使得粒子执行若干步三态量子游走, 达到相似数据点聚集到一起的目的。首先, 根据邻居集合和近邻集合构造粒子的游走空间; 然后, 根据定义的逗留概率构造了三态量子游走的硬币矩阵; 最后, 依据更新规则来实现数据点的移动。实验仿真结果证明了该算法的有效性。一维离散量子游走是最简单的量子游走模型, 还有很多复杂且有趣的量子游走模式, 如平面量子游走、散射量子游走等。在今后的研究中, 我们将对这些量子游走进行研究, 分析它们的特性, 并尝试将其应用到聚类分析中。

参考文献

- [1] Shor P W. Algorithms for quantum computation ; Discrete logarithms and factoring[C]//Proceedings of 35th Annual Symposium on Foundations of Computer Science, 1994. IEEE, 1994; 124-134
- [2] Tangwongsan S, Phoophuangpairaj R. Boosting Thai Syllable Speech Recognition Using Acoustic Models Combination[C]//International Conference on Computer and Electrical Engineering(ICCEE 2008). 2008;568-572
- [3] Gu Hung-yan, Lai Ming-uen, Tsai Sung-Feng. Combining HMM Spectrum Models and ANN Pros-ody Models for Speech Synthesis of Syllable Prom-inent Languages[C]//2010 7th International Symposium Chinese Spoken Language Processing (ISCSLP). Tainan, 2010;451-454
- [4] Thairatananond Y. Towards the Design of a Thai Text Syllable Analyzer [D]. Asian Institute of Technology, 1981
- [5] Charnyapornpong S. A Thai syllable separation algorithm [D]. Asian Institute of Technology, 1983
- [6] Poowarawan Y. Dictionary - based Thai syllable separation [C]//Proceedings of the Ninth Electronics Engineering Conference. 1986
- [7] Aroonmanakun W. Collocation and Thai Word Segmentation [C]//Proceedings of SNLP-Oriental Cocosda, 2002. 2002;68-75

- [2] Grover L K. Quantum mechanics helps in searching for a needle in a haystack[J]. Physical Review Letters, 1997, 79(2):325-328
- [3] Farhi E, Goldstone J, Gutmann S, et al. A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem [J]. Science, 2001, 292(5516):472-475
- [4] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases [J]. AI magazine, 1996, 17(3):37-54
- [5] Schuld M, Sinayskiy I, et al. An introduction to quantum machine learning[J]. Contemporary Physics, 2015, 56(2):172-185
- [6] Wang S H, Long G L. Big data and quantum computation[J]. Chinese Science Bulletin, 2015, 60(5/6):499-508(in Chinese)
王书浩, 龙桂鲁. 大数据与量子计算[J]. 科学通报(中文版), 2015, 60(5/6):499-508
- [7] Venegas-Andraca S E. Quantum walks; a comprehensive review [J]. Quantum Information Processing, 2012, 11(5):1015-1106
- [8] Childs A M. Universal computation by quantum walk[J]. Physical Review Letters, 2009, 102(18):180501
- [9] Li Q, He Y, Jiang J. A hybrid classical-quantum clustering algorithm based on quantum walks[J]. Quantum Information Processing, 2011, 10(1):13-26
- [10] Inui N, Konno N, Segawa E. One-dimensional three-state quantum walk[J]. Physical Review E, 2005, 72(5):056112
- [11] Erkan G. Language model-based document clustering using random walks[C]//Proceedings of the main conference on human language technology conference of the north American chapter of the association of computational linguistics. Association for Computational Linguistics, 2006;479-486
- [12] MacQueen J. Some methods for classification and analysis of multivariate observations[J]. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1(14):281-297
- [8] Fferty J, McCallum A, Pereira F. Conditional random fields Probabilistic models for segmenting and labeling sequence data [C]//ICML2001. San Francisco; Morgan Kaufmann, 2001;282-289
- [9] Sproat R, Emerson T. The first international Chinese word segmentation bakeoff [C]//2nd SIGHAN Workshop on Chinese Language Processing. Morristown, NJ; ACL, 2003;133-143
- [10] Zhao Hai, Huang Chang-ning, Li Mu. An improved Chinese word segmentation system with conditional random field [C]//5th SIGHAN Workshop on Chinese Language Processing. Morristown, NJ; ACL, 2006;108-117
- [11] Segmentation Guidelines for InterBEST 2009 Thai Word Segmentation; An international episode [EB/OL]. http://thailang.nectec.or.th/downloadcenter/index.php?option=com_docman&task=cat_view&gid=43&Itemid=61
- [12] Boriboon M, et al. BEST Corpus Development and Analysis [C]//International Conference on Asian Language Processing, 2009 (IALP'09). 2009;322-327

(上接第 56 页)