基于量子退火算法的大数据模糊随机挖掘方法

高 超[1] 田彦明[1]
GAO Chao TIAN Yanming

摘 要 针对当前挖掘方法在对大数据挖掘时存在挖掘可靠性低和挖掘空间聚焦能力弱的问题，引入量子退火算法，开展大数据模糊随机挖掘方法研究。利用模糊 C 均值调度算法，进行大数据模糊融合。针对非线性分布的大数据，提取大数据模糊层次聚类特征。通过量子退火算法运算，实现随机挖掘。通过对比实验证明，新的挖掘方法具备更高的挖掘可靠性，且挖掘空间聚焦能力更强，挖掘效果更理想。

关键词 量子退火算法；模糊；随机挖掘；大数据

doi：10.3969/j.issn.1672-9528.2024.08.041

0 引言

在大数据时代背景下，数据挖掘技术已经成为各领域不可或缺的分析工具。然而，随着数据规模的急剧增长和数据结构的日益复杂，现有数据挖掘方法已经难以满足高效、准确地提取有价值信息的需求。特别是当涉及模糊性和随机性的数据处理时，现有算法往往显得力不从心。目前，许多基于机器学习和深度学习的挖掘方法被广泛应用于大数据分析中。例如，殷倩倩等人 [1] 在大数据背景下，将机器学习应用到数据挖掘中；刘静等人 [2] 将数据挖掘与深度学习相结合，设计了一种全新的在线学习预测评估模型。在当前研究发展水平背景下，这些方法虽然在一定程度上提高了数据挖掘的效率和精度，但仍然存在一些固有的问题。例如，机器学习算法在处理高维度数据时容易陷入局部最优解，深度学习算法则需要大量的训练数据和计算资源 [3]。此外，这些方法对于数据的模糊性和随机性的处理能力也有限，往往难以准确捕捉数据的内在规律和模式。近年来，量子计算技术的发展为大数据挖掘提供了新的思路。量子退火算法作为一种模拟量子系统退火过程的优化算法，具有全局搜索能力强、收敛速度快等优点。因此，基于量子退火算法的大数据模糊随机挖掘方法具有广阔的应用前景。通过利用量子退火算法的全局搜索能力，可以更有效地处理大数据中的模糊性和随机性，从而提取出更多有价值的信息。本文旨在探索基于量子退火算法的大数据模糊随机挖掘方法，通过构建合适的模型和算法，实现对大规模、高维度、模糊随机数据的有效挖掘和分析。

1 大数据模糊融合

大数据的分布往往呈现出复杂的非线性特征，这使得在挖掘其有用信息时，必须深入考虑数据之间的关联性，并进行相应的融合处理。这种处理方式不仅有助于更准确地理解数据的内在结构和规律，还为后续的数据特征提取奠定了坚实的基础 [4]。为了更好地处理这种非线性分布的大数据，采用一种先进的算法——模糊 C 均值调度算法。该算法能够充分利用数据的模糊性特征，通过建立数据挖掘的语义本体模型，实现对数据的深度挖掘和精准分析。具体来说，模糊 C 均值调度算法首先对数据集进行预处理，包括数据的清洗、归一化等操作，以消除噪声和异常值对挖掘结果的影响。然后，算法根据数据的关联性进行聚类分析，将数据划分为不同的类别或簇 [5]。在聚类过程中，算法会考虑数据的模糊性，即数据点可能属于多个类别的情况，从而得到更准确的聚类结果。算法利用建立的语义本体模型对数据进行语义层面的分析和挖掘。语义本体模型的表达式为：

$$TL_X(x,y)=f(GD)_X(x,y)>q_x \quad (1)$$

式中：$TL_X(x,y)$ 表示大数据挖掘语义本体模型；$GD_X(x,y)$ 表示大数据语义协方差数值；q_x 表示阈值。通过定义和描述数据的语义概念及其之间的关系，算法能够更深入地理解数据的含义和背景，从而提取更有价值的信息 [6]。算法会对挖掘结果进行可视化展示和评估，帮助更直观地了解数据的分布和特征，以及挖掘结果的准确性和有效性。

设定大数据样本的长度为 l，然后依据其相似属性进行有序排列 [7]。为了构建大数据的模糊随机挖掘特征分布模型，采用谱分离的方式，这种方式能够有效地揭示数据的内在结构和特征。

为了进一步增强模型的准确性和实用性，利用关联语义规则对特征分布模型进行相关性融合处理。这一步骤的关键在于挖掘不同特征之间的关联性和相互作用，从而得到一个更为全面和深入的特征描述 [8]。

1. 沈阳发动机研究所 辽宁沈阳 110000

在融合处理的过程中，为了更精细地划分数据的类别和特征，采用模糊分区调度方法，对聚类中心进行特征分解。这种方法能够有效地处理数据中的模糊性和不确定性，提高聚类的准确性和稳定性[9]。随着大数据的特征分布从 l+1 维空间扩展至 2l 维空间，数据的复杂性和关联性也逐渐增加。为了有效地处理这种高维数据，采用相空间重构算法对大数据的关联特征进行重组。假设 O 表示大数据关联特征分解后得到的数据，对 O 进行盲源分离处理，得到公式：

$$O = d(O_1, O_2, \ldots, O_m)R^{m\times n} \tag{2}$$

式中：$O_1, O_2, \ldots, O_m$ 表示分离处理后的集合子集。根据前述公式计算得出的结果，进行排序处理，确保盲源分离后的大数据关联特征能够按照其数值大小进行有序排列。这一步骤的目的是更有效地实现非线性分布大数据的模糊融合。通过排序能够更加清晰地把握数据的内在规律和关联性，为后续的模糊融合提供更有力的支持。

2 大数据模糊层次聚类特征提取

在针对非线性分布的大数据进行融合处理后，为了进一步提炼其中的关键特征，采用模糊层次聚类分析算法，并结合大数据特征之间的相关性进行特征提取。这一过程不仅有助于更深入地理解数据的内在结构和规律，还为后续的数据分析和应用提供了有力的支持。具体来说，首先计算大数据的语义相似度[10]。语义相似度是衡量数据之间在意义或内容上相近程度的指标，通过计算语义相似度，可以有效地识别出具有相似特征或属性的数据点，为后续的特征提取提供重要依据。在得到语义相似度的基础上，进一步构建大数据语义关联映射微分表达式为：

$$\partial = f_j^i(k_j^i, u_i, u_j) \tag{3}$$

式中：i 表示大数据特征聚类维数；j 表示采样点数；k_j^i 表示特征聚类维数和采样点数为 i、j 的大数据样本；∂ 表示语义关联映射微分值；u_i 和 u_j 均表示特征变量。这一表达式能够直观地展示数据点之间在语义层面的关联性和变化趋势，有助于更准确地把握数据的分布和特征[11]。通过模糊层次聚类分析算法的应用，可以将数据按照其特征和属性进行层次化聚类。在聚类过程中，充分考虑数据之间的模糊性和不确定性，通过定义模糊隶属度函数来描述数据点属于不同类别的程度，从而得到更为准确和精细的聚类结果。最终，结合大数据特征的相关性，能够从聚类结果中提取关键的特征信息。这些特征不仅反映了数据的内在规律和结构，还为后续的数据分析和应用提供了重要的参考依据。

3 基于量子退火算法的随机挖掘

基于量子退火算法的随机挖掘是一种结合量子计算和随机搜索技术的数据挖掘方法。该方法利用量子比特的量子叠加和量子纠缠特性，在解空间中搜索最优解，以实现高效地随机挖掘。

量子退火算法的基本原理是通过量子比特的量子叠加和量子纠缠来搜索解空间[12]。在量子计算中，量子比特可以处于多个状态的叠加态中，这使得量子计算机可以同时处理多个可能的解。量子纠缠则是一种量子比特之间的相互作用，它可以使量子计算机在搜索解空间时更加高效。

基于量子退火算法的随机挖掘过程包括以下几个关键步骤。

第一步，初始化：根据具体问题，设定初始的权重和候选状态，构建相应的物理量子叠加态$|\mathrm{Y}\rangle$，表达式为：

$$|\mathrm{Y}\rangle = a|0\rangle + b|1\rangle \tag{4}$$

式中：$|0\rangle$ 和 $|1\rangle$ 表示量子比特的基态；a 和 b 表示复数，满足 $a|^2 + |b|^2 = 1$ 的条件。

第二步，量子演化：按照含时薛定谔方程开始量子演化，利用量子比特的叠加和纠缠特性，在解空间中搜索可能的解。量子的状态随时间的变化由含时薛定谔方程描述为：

$$i\frac{\mathrm{d}}{\mathrm{d}t}|Y(t)\rangle = H(t)|Y\rangle \tag{5}$$

式中：$|\mathrm{Y}(t)\rangle$ 表示随时间变化的量子态；$H(t)$ 表示含时的哈密顿量。

第三步，量子穿隧：根据横向场的时间依赖强度，在不同的状态之间产生量子穿隧，使候选状态不断改变，实现量子并行性。

第四步，结果提取：当横向场最终被关闭时，预期达到原优化问题的解，即到达相对应的经典伊辛模型基态。量子退火算法通过逐渐改变退火参数 s 从 0 到 1，从初始哈密顿量的基态逐渐过渡到稳态哈密顿量的基态。这个过程通常是一个缓慢变化的过程，以确保能够找到全局最优解。

4 对比实验

4.1 实验准备

为了全面验证本文设计的基于量子退火算法的挖掘方法在实际场景中的实用性及有效性，搭建了基于 MATLAB 软件和 VC++ 编译环境的实验框架。选取某高校网络用户的使用数据作为实验的核心数据集，其中涵盖高达 50 000 个大数据样本，这些样本具有丰富的语义信息和实际应用价值。为了确保实验的准确性和可靠性，根据特定的语义关联规则，将这些数据精心划分为 400 个训练集。

在实验过程中，特别设定大数据的采集频率为 1 kHz，这一频率既能保证数据的实时性，又能兼顾数据处理的效率。同时，为了更好地处理数据中的模糊性，设定 35 个关联语义集合，用于区分和识别不同数据之间的模糊关系。

为了全面评估本文提出的挖掘方法的性能，设置了两组对照实验。对照 A 组采用当前流行的基于无监督深度学习的挖掘方法，而对照 B 组则使用经过改进的模糊聚类算法进行挖掘。通过对比实验组与对照 A 组、对照 B 组的挖掘结果，

能够更加客观、全面地评估基于量子退火算法的挖掘方法在实际应用中的表现。

图 1 展示了实验中使用的 5000 个大数据样本的采集分布情况。

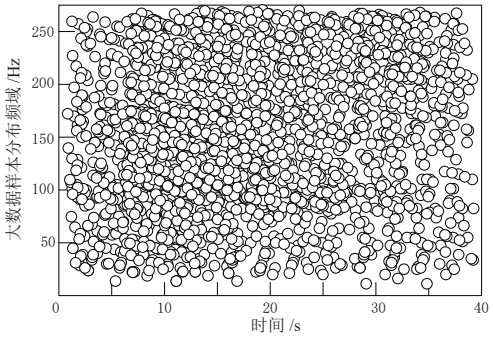

图 1 5000 个大数据样本采集分布示意图

从图 1 中可以清晰地看到，这些数据样本在时间上呈现出一定的分布规律，且不同样本之间的语义关联也错综复杂。

4.2 三种方法挖掘可靠性分析

在数据挖掘的过程中，若出现欠拟合情况，则说明数据挖掘结果不准确。欠拟合意味着模型未能充分捕捉数据中的内在规律和模式，导致挖掘结果偏离真实情况，进而影响到最终决策的有效性。因此，将欠拟合情况作为挖掘可靠性的重要评价指标，对于提升数据挖掘的准确性和实用性至关重要。将上述大数据集合作为测试样本，分别利用三种挖掘方法对大数据进行挖掘，并结合 MATLAB 将三种挖掘方法的拟合曲线与数据点分布情况进行绘制。若绘制结果中拟合曲线变化与数据分布一致，则说明没有出现过拟合情况，挖掘可靠；反之，若拟合曲线变化与数据分布不一致，则说明出现过拟合情况，挖掘不可靠。图 2~ 图 4 为三种挖掘方法的可靠性实验结果。

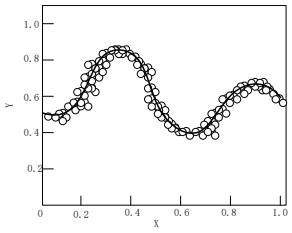

图 2 实验组大数据挖掘可靠性实验结果图

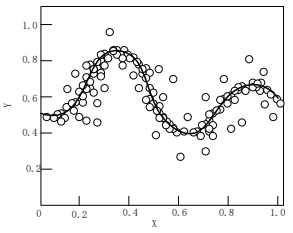

图 3 对照 A 组大数据挖掘可靠性实验结果图

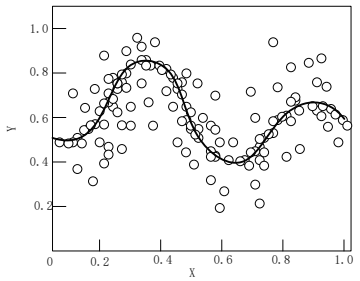

图 4 对照 B 组大数据挖掘可靠性实验结果图

通过对图 2~ 图 4 的深入对比分析，可以清晰地看到不同挖掘方法在拟合曲线与数据分布之间的表现差异。首先，观察实验组的拟合曲线，可以发现其变化与数据分布高度一致，几乎每一个数据点都紧密地贴合在曲线上，呈现出一种统一的态势。这种一致性不仅表明实验组的方法能够准确地捕捉数据的内在规律，还反映出其强大的泛化能力和稳定性。然而，当将视线转向对照 A 组和对照 B 组时，情况则大不相同。在这两组实验结果中，明显看到拟合曲线与数据分布之间存在较大的偏差。对照 A 组的拟合曲线在某些区域虽然能够较好地拟合数据，但在其他区域则出现了明显的偏离，部分数据点甚至分布在距离拟合曲线较远的位置。这种不一致性说明对照 A 组的方法在处理数据时可能存在一定的局限性，无法全面、准确地反映数据的真实情况。对照 B 组的情况则更为严重。其拟合曲线与数据分布的偏差更为明显，大量数据点远离拟合曲线，分布得相当散乱。这种明显的偏差不仅表明对照 B 组的方法在拟合数据方面存在严重问题，还暗示其可能无法有效处理数据的模糊性和随机性。综合以上分析，可以得出结论：实验组基于量子退火算法的挖掘方法在拟合曲线与数据分布的一致性方面表现最佳，其挖掘结果的可靠性最强。相比之下，对照 A 组和对照 B 组的方法在处理大数据挖掘任务时存在一定的不足和局限性。因此，在实际应用中，可以优先考虑采用实验组的方法来提升数据挖掘的准确性和可靠性。

4.3 三种方法挖掘空间聚焦能力对比

对三种挖掘方法的空间聚焦能力进行对比，将上述大数据样本集合中的 20 000 个数据作为训练样本，使用三种挖掘方法对该大数据进行挖掘，并给出挖掘前后大数据实部和虚部的分布情况。图 5 为挖掘前大数据样本分布情况示意图。

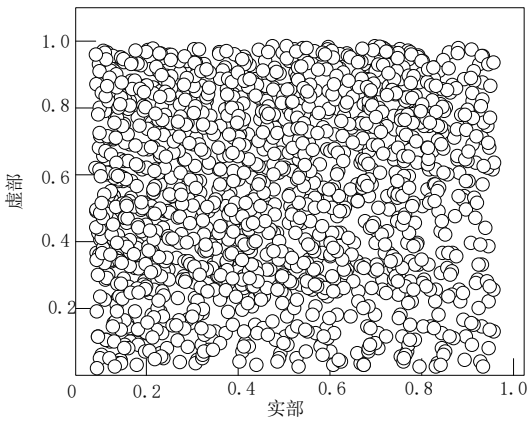

图 5 挖掘前大数据样本分布情况示意图

从图 5 可以看出，大数据样本分布的实部和虚部均在 -1 ～ 1 范围内，呈整片分布状态，且存在大量数据散点。将经过三种挖掘方法挖掘后的结果绘制成图 6 ～图 8 所示。

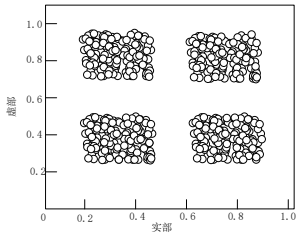

图 6 实验组大数据挖掘可靠性实验结果图

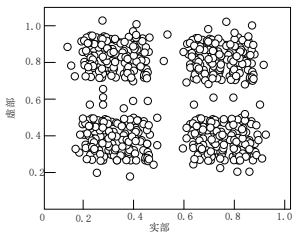

图 7 对照 A 组挖掘空间聚焦能力对比图

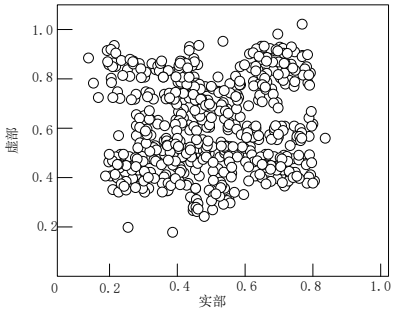

图 8 对照 B 组挖掘空间聚焦能力对比图

通过仔细对比图 6 ～图 8 所展示的实验结果，可以明显观察到不同挖掘方法在处理大数据样本时的空间聚焦能力差异。在实验组的结果中，可以看到经过基于量子退火算法的挖掘方法处理后，数据呈现出一种高度规律的状态分布。不同特征的数据紧密地聚集在一起，形成了清晰的聚类，数据分布边界明显，易于区分。这种规律性的分布情况不仅反映了数据之间的内在关联，也展示了挖掘方法强大的空间聚焦能力。相比之下，对照 A 组和对照 B 组的结果则显得杂乱无章。对照 A 组虽然能够大致看出四个分区，但数据分布并不规律，不同特征之间的数据边界模糊，难以准确划分。而对照 B 组的结果更是让人失望，数据分布完全没有规律可言，不同特征的数据混杂在一起，无法实现有效的空间聚焦。这种明显的差异不仅体现在数据的整体分布上，还体现在每一个细节之处。在实验组的结果中，可以看到每一个聚类都是紧凑而清晰的，数据点之间的距离适中，没有出现过于密集或过于稀疏的情况。而对照 A 组和对照 B 组的结果中，数据点之间的距离则显得杂乱无章，无法形成有效的聚类。综上所述，通过对比图 6~ 图 8 的实验结果，可以得出结论：实验组基于量子退火算法的挖掘方法在空间聚焦能力方面表现最为出色。该方法不仅能够准确捕捉数据之间的内在关联，还能够实现有效的数据聚类，为后续的数据分析和决策提供有力支持。因此，在实际应用中，可以优先考虑采用实验组的方法来提升数据挖掘的空间聚焦能力。

5 结语

本研究通过深入探索基于量子退火算法的大数据模糊随机挖掘方法，不仅为解决当前数据挖掘领域的挑战提供新的思路，也为未来的大数据应用开辟了新的可能性。尽管量子计算技术目前仍处于发展阶段，但其强大的计算能力和独特的优化机制已经展现出了巨大的潜力。本文所提出的方法在理论和实验层面均取得了一定成果，不仅提高了数据挖掘的效率和精度，还增强了对模糊随机数据的处理能力。然而也意识到，这一方法在实际应用中仍面临诸多挑战，如量子计算资源的有限性、算法的稳定性和可扩展性等。展望未来，将继续深化对量子退火算法的研究，优化算法参数和模型结构，以进一步提升其在大数据挖掘中的应用效果。

参考文献：

[1] 殷倩倩 , 申鑫欣 , 夏祎 . 大数据背景下机器学习在数据挖掘中的应用 [J]. 数字技术与应用 ,2022,40(5):21-23.

[2] 刘静 , 由志勋 . 基于深度学习与数据挖掘的在线学习预测评估模型设计 [J]. 电子设计工程 ,2023,31(15):131-134+139.

[3] 赵星 . 大数据支持下的农机作业数据挖掘与决策分析技术研究 [J]. 南方农机 , 2024, 55(5):182-184.

[4] 谢莉 . 基于大数据分析的通信网络部门档案价值挖掘与应用 [J]. 办公室业务 , 2024(5):90-92.

[5] 姚丽敏 , 马允雪 . 大数据角度下以数据挖掘为支持的科研管理系统设计思考 [J]. 山西能源学院学报 ,2024,37(1):56-58.

[6] 顾耀文 , 李姣 . 基于无监督深度学习的电子健康档案数据挖掘技术研究进展 [J]. 医学信息学杂志 ,2022,43(1):34-40.

[7] 岳宝强 , 杨波 , 李彪 , 等 . 基于数据挖掘和 LSSVM 的电量大数据多维感知方法 [J]. 微型电脑应用 ,2023,39(12):58-61+84.

[8] 李萍 , 刘金金 . 基于改进模糊聚类算法的大数据随机挖掘仿真 [J]. 计算机仿真 ,2024,41(2):496-499+521.

[9] 万超 . 大数据挖掘在供电所台区线损管理工作中的应用实践研究——以东马坊供电所为例 [J]. 电子元器件与信息技术 , 2023,7(11):113-116.

[10] 陈晓姗 , 张国华 . 基于朴素贝叶斯的大数据模糊随机挖掘仿真 [J]. 计算机仿真 , 2023,40(11):428-432.

[11] 刁军飞 , 杨帆 , 温羽 , 等 . 基于大数据挖掘中医治疗围绝经期潮热组方配伍规律的研究 [J]. 海峡药学 , 2023, 35(11): 26-30.

[12] 郭宇 , 张传洋 , 刘梦婷 , 等 . 用户信息焦虑纾解驱动医疗健康大数据价值实现影响因素与路径研究 [J]. 图书情报工作 , 2023,67(15):25-34.

【作者简介】

高超（1984—），男，辽宁昌图人，本科，高级工程师，研究方向：网络安全、数据安全、航空信息化。

田彦明（1978—），男，辽宁兴城人，本科，高级工程师，研究方向：航空信息化。

（投稿日期：2024-05-14）