文章编号:1671-4598(2026)03-0194-07 DOI:10.16526/j.cnki.11-4762/tp.2026.03.024 中图分类号:TP393 文献标识码:A

基于云计算技术的海量云数据模糊聚类算法设计

罗 萍,张 雷
(中国民用航空飞行学院 信息中心,成都 610000)

摘要:云数据呈现出爆炸式增长,其规模海量、来源多元异构、结构复杂且动态变化显著;为实现高维、复杂云数据的高效处理、增强对云数据不确定性和模糊性的适应能力,设计基于云计算技术的海量云数据模糊聚类算法;构建基于云计算的海量云数据分析框架,主节点服务器采用随机森林算法实现来自多个异构源的海量云数据融合后,在对其作切分处理后,将得到的多个云数据切片分配给从节点服务器,计算节点在MapReduce数据模型下调用模糊K-means算法执行本地云数据聚类任务,采用量子粒子群算法优化初始聚类中心后,输出云数据聚类结果;实验结果表明:该方法可实现云数据模糊聚类,簇内云数据呈现紧凑分布形态,簇间数据区分度高;聚类中心优化选择后,聚类误差降低至0.10左右,分离系数为0.891,分离熵为10.441;计算节点数量为10时,加速比达到最大。

关键词:云计算;模糊聚类;随机森林;模糊K-means算法;MapReduce;量子粒子群

Design of Fuzzy Clustering Algorithm for Massive Cloud Data Based on Cloud Computing Technology

LUO Ping, ZHANG Lei
(Centre of Information, Civil Aviation Flight University of China, Chengdu 610000, China)

Abstract: With the explosive growth of cloud data, it has the characteristics of massive scale, diverse and heterogeneous sources, complex structure, and significant dynamic changes. To achieve efficient processing for high-dimensional and complex cloud data and enhance adaptability to the uncertainty and fuzziness of cloud data, a fuzzy clustering algorithm for massive cloud data based on cloud computing technology is designed, which constructs a cloud computing-based framework to analyze massive cloud data. The master node server employs the random forest algorithm to fuse massive cloud data from multiple heterogeneous sources. After slicing the data, the resulting multiple cloud data slices are distributed to slave node servers. Based on the MapReduce data model, computing nodes invoke the fuzzy K-means algorithm to perform local cloud data clustering tasks. After the quantum particle swarm algorithm is used to optimize the center of initial clusters, the cloud data clustering results are output. Experimental results indicate that this method can achieve the fuzzy clustering of cloud data, with cloud data within clusters exhibiting a compact distribution and high differentiation between clusters. After optimizing the selection of cluster centers, the clustering error is reduced to around 0.10, with a separation coefficient of 0.891 and a separation entropy of 10.441. The speedup ratio reaches its maximum with the number of computing nodes being 10.

Keywords: cloud computing; fuzzy clustering; random forest; fuzzy K-means algorithm; MapReduce; quantum particle swarm

0 引言

在当今数字化时代,云数据呈现出规模海量、类型多样、动态变化等显著特点[1],蕴含的丰富、潜在价值隐藏在复杂的数据结构与模糊的类别边界之中。传统聚类算法难以应对云数据中广泛存在的模糊性与不确定性[2-3]。模糊聚类凭借其允许数据对象以不同程度隶属于多个类别的特性,能够更贴合实际地刻画数据的内在分布规律,为深入挖掘云数据中的潜在模式与关联信息提供了有效途径[4-6]。对海量云数据进行模糊聚类不仅

收稿日期:2025-08-01; **修回日期**:2025-10-14。
作者简介:罗 萍(1991-),女,大学本科,工程师。
引用格式:罗 萍,张 雷.基于云计算技术的海量云数据模糊聚类算法设计[J].计算机测量与控制,2026,34(3):194-200.

有助于揭示数据背后的复杂关系，更能为精准决策、个性化推荐、风险评估等应用提供有力支持。

张文宇等[7]利用天牛群优化算法对 DBSCAN 算法中的关键参数进行优化选择后，实现数据的聚类挖掘。该算法在面对密度不均匀数据时，可能无法有效区分簇，导致聚类效果不理想。李艳霞等[8]利用 K-means 聚类算法处理物联网数据，通过粒子群算法进行初始聚类中心的优化选择，提高数据聚类效果。然而，该算法在边界模糊数据集处理上存在局限性。张嘉旭[9]等结合核范数、香农熵理论与模糊聚类算法处理多视觉数据，实现一致性与差异性数据的有效区分。该算法在处理大规模数据集时，参数优化过程需要大量的迭代和计算资源，可能导致算法的运行效率较低。

云计算技术[10-11]以其强大的计算能力、灵活的资源调配以及高效的并行处理机制，为海量云数据聚类带来了新的机遇，通过将模糊聚类算法部署于云计算平台，能够充分利用其分布式计算架构，有效解决传统聚类算法在处理大规模数据时面临的计算瓶颈与存储限制。因此，本文设计基于云计算技术的海量云数据模糊聚类算法，显著提升聚类效率与可扩展性，使得海量云数据的模糊聚类分析得以高效、稳定地开展，进而推动数据挖掘与智能分析技术在各个领域的广泛应用与发展。

1　云数据模糊聚类

1.1　基于云计算技术的海量云数据分析框架

本文依托云计算平台[12-14]，运用主节点与从节点（Master-Slave）的架构模式，对大规模云数据进行分散式存储和协同计算。通过运用模糊聚类算法处理海量云数据，深入挖掘数据背后的潜在价值，实现海量云数据的精准分类。基于云计算的海量云数据分析架构如图 1 所示。

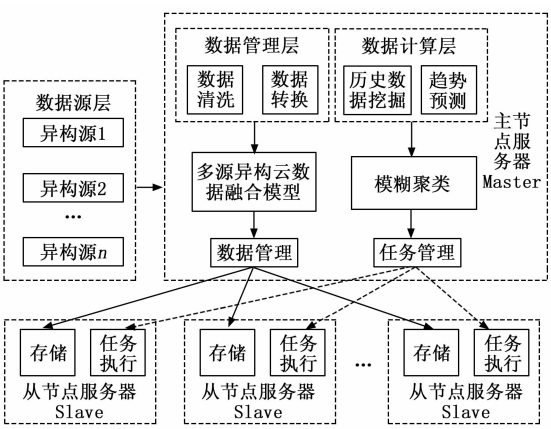

图 1　基于云计算的海量云数据分析架构

在该架构中，主节点服务器占据核心地位，承担着海量云数据管理与计算任务分配的关键职责，能够对来自多个异构数据源的海量云数据进行全面整合处理，历经数据转换、清洗等精细化操作流程，构建出多源异构云数据融合模型。这一模型为后续的数据分析与深度挖掘提供了坚实的数据基础，确保数据分析的准确性与全面性。与此同时，主节点服务器还肩负着协调计算层作业的重要使命。它运用先进的模糊聚类算法，对历史云数据进行深度剖析与解读，高效构建数据挖掘模型。通过对数据的精准建模，能够实现对云数据的精细化分析，为业务决策提供科学依据。

1.2　海量多源异构云数据的融合

1.2 节在 1.1 节搭建的云计算分析框架下，运用随机森林算法实现海量云数据的融合。随机森林算法的应用是在主节点服务器的管理和协调下进行的，通过对多个异构源的云数据子集进行处理，实现数据格式统一、构建融合模型等操作，是 1.1 节所构建框架下的具体算法实现环节，进一步细化和落实了整体的数据处理流程，为实现 1.1 节中设定的数据分析目标提供了整合后的高质量数据。

来自多异构源的海量云数据集合 X 由 n 个云数据子集 x_1，x_2，…，x_n 构成，描述公式为：

$$X=\{f(x_1),f(x_2),\cdots,f(x_n)\} \tag{1}$$

其中：利用映射推理函数 f 对每个云数据子集 x_i 进行处理，实现数据格式统一。为每个云数据子集设定的处理任务表示为 g_1，g_2，…，g_n，通过下式描述多源异构云数据集合：

$$D=\{(x_1,g_1),(x_2,g_2),\cdots,(x_n,g_n)\} \tag{2}$$

在实际的海量多源异构云数据中，存在大规模缺乏标签的数据。这类无标签数据的存在，极大地增加了数据融合的复杂性和难度，因为缺乏明确的类别标识，使数据之间的关联和整合缺乏有效的引导信息。因此，本文在 1.1 小节的云计算分析框架下，通过随机森林算法实现 D 中海量云数据的融合。随机森林中包含 50 棵 CART（Classification and Regression Trees）决策树。选择 50 棵决策树是基于多方面的考量，一方面，较多的决策树数量能够增强模型的稳定性和泛化能力，减少因个别决策树的误差对整体融合效果的影响；另一方面，经过实验和经验验证，50 棵决策树在保证融合效果的同时，不会使计算复杂度过高，能够在合理的时间内完成数据融合任务。

在特征选择策略上，采用 Gini 系数作为选择分裂特征的依据。Gini 系数是一种衡量数据不纯度的指标，通过计算不同特征划分数据后的 Gini 系数，选择使 Gini 系数下降最快的特征作为分裂特征。这种方式能够有效地选择出对数据分类或区分能力较强的特征，从而提高决策树的分裂质量，进而提升数据融合的效果。

对于每棵决策树的训练样本选取，采用随机选取60%样本的方式进行训练。这样做的好处在于，每棵决策树基于不同的样本子集进行训练，增加了决策树之间的差异性，使得随机森林模型能够捕捉到数据中更多不同的模式和特征，进一步提高了模型的泛化能力和对复杂数据的适应能力，最终实现更优的海量多源异构云数据融合效果。

决策树 h_i 通过下式进行构建：

$$h_i = h(x_i, \theta_i) \tag{3}$$

其中：h_i 构建过程中涉及的所有参数表示为 θ_i。通过下式确定海量多源异构云数据融合模型 H：

$$H = \{h_1, h_2, \cdots, h_i, \cdots, h_n\} \tag{4}$$

$ep(p \mid x_i)$ 表示 x_i 中各类别云数据的分类效果评价概率，可通过随机森林中各决策树的预测结果统计得出。具体而言，对融合后的云数据，每棵决策树基于其训练样本生成类别预测，最终融合模型通过投票或概率平均机制汇总所有决策树的预测结果。例如，若某样本被 50 棵决策树中的 30 棵归类为类别 A，则其属于 A 类的分类效果评价概率为 30/50=0.6。该概率值用于评估融合模型对各类别数据的区分能力，值越高表明模型对该类别的分类效果越优，进而验证数据融合后是否满足后续分析的准确性要求。根据其可计算出 X 的分类效果评价概率，从而衡量海量多源异构云数据融合效果，计算公式为：

$$ep(p \mid X) = n^{-1} \sum_{i=1}^{n} ep(p \mid x_i) \tag{5}$$

如果融合后的云数据集 H 在分类任务上表现出与真实情况相符的趋势，需满足下式边界条件：

$$\zeta(x_i, p) > 0 \tag{6}$$

基于随机森林的海量多源异构云数据融合达到预期效果，需使泛化偏差 η 满足下式条件：

$$\eta = E[m(x_i, p)] < 0 \tag{7}$$

其中：云数据分类效果评价期望值表示为 $E[m(x_i, p)]$。

1.3 基于模糊 K-means 的海量云数据聚类

云数据因其高维性、复杂性以及动态变化性，使得其类别边界往往呈现出模糊且连续渐变的特征[15]。传统的聚类算法在处理此类数据时，难以准确刻画数据样本在不同类别间的复杂归属关系。而模糊 K-means 聚类算法在经典 K-Means 算法的基础上创新性地引入了隶属度概念，这一改进使其能够以一种更为灵活的方式处理云数据样本在各类别间的归属问题，进而更精准地挖掘云数据内在的结构和分布规律[16-17]。因此，本文采用模糊 K-means 算法处理融合后的云数据集 H。设定聚类数目表示为 K，对 H 进行模糊 K-means 聚类后，可得到 i 个类别的云数据子集 x_1，x_2，…，x_i，与之对应的聚类中心表示为 c_1，c_2，…，c_i，且具备下列条件：

1）每个聚类子集非空，即 $x_i \neq \varnothing$。

2）不同聚类集合之间互不相交，即 $x_i \cap x_j \neq \varnothing$。

3）所有聚类子集的并集等于融合后的云数据集 H，即 $\bigcup_{i=1} x_i = H$。

模糊 K-means 聚类算法以每个云数据点至所属聚类中心欧几里得距离之和最短为优化目标，计算公式描述为：

$$J(W, C) = \min \sum_{i=1}^{n} w_{ij}^{m} \sum_{k=1}^{n} \sum_{j=1}^{K} \| h_{ik} - c_{jk} \| \tag{8}$$

式中，m 表示模糊权重系数；隶属度参数 w_{ij} 反映了云数据 h_i 相对于类别 C_j 的归属程度，$0 \leqslant w_{ij} \leqslant 1$，且需使下式成立。

$$\sum_{j=1}^{K} w_{ij} = 1 (1 \leqslant i \leqslant n) \tag{9}$$

通过拉格朗日乘子法确定 w_{ij} 和 c_j，计算公式描述为：

$$w_{ij} = \begin{cases} \dfrac{1}{\sum_{q=1}^{K} \left(\dfrac{\| h_i - c_j \|}{\| h_i - c_q \|} \right)^{2/m-1}} \\ 1, \| h_i - c_q \| = 0 (i = q) \\ 0, \| h_i - c_q \| = 0 (i = q) \\ 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant K \end{cases} \tag{10}$$

$$c_j = \sum_{i=1}^{n} w_{ij}^{m} x_i / \sum_{i=1}^{n} w_{ij}^{m}, 1 \leqslant j \leqslant K \tag{11}$$

1.4 基于量子粒子群算法的聚类中心确定

聚类中心的选择是决定海量云数据聚类效果的关键因素[18-19]。本文通过量子粒子群算法实现聚类中心的优化选择，以实现海量云数据的精准分类。种群个体代表了一个潜在的解（聚类方案），包含了 K 个聚类中心。由于每个云数据的维度表示为 B，故种群个体可通过一个 $K \times B$ 维向量描述。第 l 个个体的位置通过下式描述：

$$Z_l = (c_{l1}, c_{l2}, \cdots, c_{li}, \cdots, c_{lk}) \tag{12}$$

式中，该个体的第 i 个聚类中心表示为 c_{li}。基于量子粒子群算法的聚类中心优化选择步骤如下。

第一步：对参数 K、B、种群大小 N 以及迭代轮次上限 T 进行初始设置。

第二步：采用模糊 K-means 聚类算法[20-21]处理融合后的云数据样本集合，对得到的初始聚类中心进行编码，生成种群个体 Z_1。根据欧式距离进行一次初步聚类操作，将云数据划分成 K 个类别子集 S_1，S_2，…，S_K。

第三步：从每个类别子集任选一个云数据，通过编码得到种群个体 Z_2。反复执行该过程，直到产生全部

种群个体为止。

第四步：分别计算种群个体的局部、全局最佳位置，表示为 p_{best}、p_{gbest}。

第五步：将公式（8）作为种群个体适应度函数，确定每个种群个体的适应度值后，通过下式对 p_{best}、p_{gbest} 进行动态修正：

$$p_{\text{best}}^{(t)} = \begin{cases} z_{id}, f(x_{id}^{(t)}) \geqslant f(p_{\text{best}}^{(t-1)}) \\ p_{\text{best}}^{(t-1)}, f(x_{id}^{(t)}) < f(p_{\text{best}}^{(t)}) \end{cases} \tag{13}$$

$$p_{\text{gbest}} = \min_{i=1,\cdots,N} p_{\text{best}}^{(i)} \tag{14}$$

通过下式计算种群中所有个体最优位置的均值：

$$mbest = \frac{\sum_{i=1}^{N} p_i(t)}{N} \tag{15}$$

第六步：利用下式生成新的种群个体：

$$z_{id} = p_{id}(t) \pm \alpha \mid p_{\text{best}}(t) - mbest \mid \times \ln(u^{-1}) \tag{16}$$

$$p_{id} = \varphi \times p_{id}(t) + (1-\varphi) p_{\text{gbest}}(t) \tag{17}$$

$$\alpha = \frac{(\alpha_1 - \alpha_2)(T-t)}{T} + \alpha_2 \tag{18}$$

式中，参数 α 控制着种群个体在搜索空间中的探索和开发能力，其初值表示为 α_1，其最终值表示为 α_2；当前迭代轮次表示为 t；φ、p_{id} 分别为平衡系数以及任意加权点，u、φ 的取值区间均为（0，1）。

第七步：当不等式 $\| J_m^{(k+1)} - J_m^{(k)} \| < \varepsilon$ 成立或达到迭代轮次上限后，停止迭代；反之，回到第四步。

第八步：输出全局最优解，确定与之对应的聚类中心，即可实现模糊 K-means 聚类算法聚类中心的优化选择。

1.5　基于 MapReduce 的海量云数据模糊聚类并行化

为了提高海量云数据模糊聚类效率，在基于云计算的海量云数据分析架构下，将融合后的海量云数据进行切分，并分配到不同从节点上，利用 MapRedue 数据模型实现多个从节点上云数据模糊聚类的并行化处理，具体实现流程如图 2 所示。基于 MapReduce 的海量云数据模糊聚类并行化处理共分为 3 个阶段，具体如下。

1）Map 阶段数据生成与初步处理：

对融合后的云数据进行切分，得到多个云数据切片后，针对每个云数据，确定其与所有聚类中心的欧几里得距离，并根据模糊 K-means 聚类算法确定隶属度 w_{ij}，从而生成包含云数据以及对应聚类中心、隶属度的中间键值对 $[K,(h_i,w_{ij})]$。同时，根据 w_{ij} 计算局部适应度值 $J_i^{(t+1)}$，生成包含标签与 $J_i^{(t+1)}$ 的键值对。标签用于标识对应的云数据或计算任务，局部适应度值则反映了在当前局部计算中的性能指标。

2）数据分块与 Shuffle 阶段传输：

每个 Map 任务生成的中间键值对按照 128 MB 的块大小进行划分。将划分后的数据块通过 Shuffle 阶段传输至 Reduce 节点。在 Shuffle 阶段，采用 TCP/IP 协议进行数据传输。为了优化传输效率，对 TCP/IP 的块大小进行合理设置，根据网络环境和数据特性，将其设置为 64 kB。这样可以减少网络传输中的报头开销，提高数据传输的效率和稳定性。

在传输过程中，系统会根据键值对中的键（即“聚类中心 ID-数据索引”）对数据进行分类和路由，确保相同聚类中心相关的数据能够准确地传输到对应的 Reduce 节点。例如，所有键中包含“C1”的键值对会被传输到负责处理聚类中心“C1”相关计算的 Reduce 节点。

3）Reduce 阶段数据接收与处理：

Reduce 节点接收到从多个 Map 节点传输过来的数据块后，首先对数据进行解析和整合。根据键值对中的键，将属于同一聚类中心的数据进行归类。然后，结合从不同 Map 节点接收到的局部适应度值等信息，完成局部模糊聚类。

在完成局部模糊聚类后，Reduce 节点对结果进行汇总，并重新确定各类别聚类中心。通过这种方式，实现了海量云数据在不同节点间的有效通信和协同计算，最终完成海量云数据的模糊聚类。

通过以上详细的数据通信流程，确保了 MapReduce 模型在海量云数据模糊聚类并行化处理中的高效性和准确性，充分利用了分布式计算资源，提高了数据处理的速度和质量。

2　实验分析

以某云计算平台中的海量云数据为实验对象，该平台中部署了一台主节点服务器（MM）、十台从节点服务器（NM1-NM10）以及一台 500 Mbps 数据交换机（MS）。主从节点服务器的硬件配置信息如表 1 所示。对云平台中的海量多源异构云数据进行采集，构建实验数据集，数据总量达到 882 680 条，涉及文本数据、数值数据、日期时间数据以及图像数据、音频数据等多种类型。数据维度分布情况如下：

表 1　主从节点服务器硬件配置信息

硬件	主节点服务器	从节点服务器
CPU	Intel Xeon E5-2690 v4	Intel Xeon E5-2660 v4
内存	128 GB DDR4	64 GB DDR4
硬盘	2 TB SAS HDD	1 TB SAS HDD
最高主频	3.6 GHz	3.2 GHz
RAM	32 GB	16 GB
ROM	512 MB	256 MB

文本数据：文本数据的维度主要取决于其包含的特征数量。在本数据集中，文本数据的维度范围在 50～

500 之间。这是因为不同的文本数据来源和内容导致其特征数量有所差异。例如，一些简短的描述性文本可能仅包含 50 个左右的特征，而较为复杂的报告或文章类文本可能具有多达 500 个特征。这些特征可能包括词汇频率、语法结构、语义信息等多个方面。

数值数据：数值数据的维度相对较为灵活，根据不同的测量指标和采集需求而定。在本数据集中，数值数据的维度范围在 10～100 之间。例如，一些基本的物理量测量数据可能仅包含 10 个左右的维度，而复杂的统计数据或金融数据可能具有多达 100 个维度。

日期时间数据：日期时间数据通常以特定的格式呈现，其维度相对固定。在本数据集中，日期时间数据一般以年、月、日、时、分、秒等形式表示，维度为 6。

图像数据：图像数据的分辨率对数据的维度有着重要影响。本数据集中的图像数据分辨率为 256×256 像素。对于彩色图像，每个像素点通常包含红、绿、蓝 3 个通道的信息，因此一张图像的数据维度为 256×256×3＝196 608。

音频数据：音频数据的维度取决于采样率、量化位数和音频时长等因素。在本数据集中，音频数据经过预处理后，其维度范围在 1 000～10 000 之间。

特征相关性如下：

各类数据特征之间的相关性系数在 0.2～0.6 之间。这表明不同特征之间存在一定的关联，但并非高度相关。例如，在文本数据中，某些词汇的出现频率可能与文本的主题相关，但并不是绝对的线性关系；在数值数据中，不同的测量指标之间可能存在一定的统计关联，但也受到多种因素的影响。这种适度的特征相关性既保证了数据具有一定的结构和规律可循，又增加了数据处理的复杂性和挑战性。

按照 1∶3∶6 比例对实验数据集进行随机分组，得到三个不同规模的云数据集，分别标记为组 1、组 2、组 3，应用云计算模糊聚类算法对 3 组海量云数据进行模糊聚类，分析其在云数据分类中的性能优势。

在量子粒子群算法中，平衡系数和任意加权点的取值对聚类中心的选择以及最终的聚类效果有着重要影响。这两个参数的取值范围均为（0，1），为了确定最优参数组合，进行了多组实验，对比不同参数组合下的聚类误差。结果如表 2 所示。

从上述实验结果可以看出，当平衡系数＝0.3，任意加权点＝0.7 时，聚类误差达到最小值 0.215，在收敛速度和避免陷入局部最优之间取得了较好的平衡。粒子既能够充分利用群体信息，又保留了一定的自身认知，从而能够更有效地搜索到合适的聚类中心，降低了聚类误差。因此，选择平衡系数＝0.3，任意加权点＝0.7 作为最优参数组合应用于量子粒子群算法中，以实现海量云数据聚类中心的优化选择。

表 2 量子粒子群算法参数敏感性分析及最优参数确定

实验编号	平衡系数	任意加权点	聚类误差
1	0.1	0.2	0.258
2	0.1	0.5	0.242
3	0.1	0.8	0.265
4	0.3	0.1	0.236
5	0.3	0.3	0.221
6	0.3	0.7	0.215
7	0.5	0.2	0.233
8	0.5	0.5	0.228
9	0.5	0.9	0.240
10	0.7	0.3	0.230
11	0.7	0.6	0.225
12	0.7	0.9	0.238

在本文所采用的模糊 K-means 算法中，模糊权重系数（与模糊隶属度相关）设定为 2。这一数值的选取是基于多方面的考量，具体如下。

平衡模糊性与确定性：模糊权重系数决定了聚类过程中的模糊程度。设置为 2 是一种在众多实际应用中较为常用的取值，它能够在数据的模糊性和确定性之间达到一种平衡。如果该系数取值过低，算法可能会趋向于产生类似硬聚类的结果，失去处理数据模糊性的优势；反之，如果取值过高，聚类结果可能会过于模糊，难以明确数据点的归属。

适应云数据特性：考虑到云数据的高维性、复杂性和动态变化性，其类别边界往往模糊且连续渐变。模糊权重系数为 2 能够使算法更好地适应这种特性，更灵活地处理云数据样本在不同类别间的归属关系，从而更精准地挖掘云数据内在的结构和分布规律。

为研究聚类中心对云数据模糊聚类效果的影响，应用云计算模糊聚类算法对第三组云数据集进行模糊聚类，聚类中心优化选择前后聚类误差差异如图 2 所示。

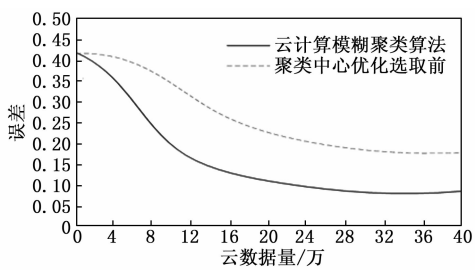

图 2 聚类中心对云数据模糊聚类结果影响分析

分析图 2 得出，随着云数据量的不断增大，聚类误差曲线呈现出持续下降的走势规律。这表明模糊 K-means 聚类算法能够有效学习和捕捉云数据中的分布特

征，有利于海量云数据的准确分类。这是因为模糊 K-means 聚类算法在处理更多数据时，能够更全面地学习和捕捉云数据中的分布特征。算法通过不断迭代调整聚类中心，使得每个数据点到其所属聚类中心的距离之和逐渐减小，从而降低了聚类误差。

利用分离系数、分离熵指标评价云数据模糊聚类效果，为了分析云计算模糊聚类算法在云数据模糊聚类上的性能优势，将其与基于改进天牛群优化的 DBSCAN 聚类算法、基于改进粒子群优化的 K-means 聚类算法进行对比，3 种算法在这两个指标上的表现差异如表 3 所示。

表 3　不同算法下的云数据模糊聚类性能对比

算法	分离系数	分离熵
云计算模糊聚类算法	0.891	10.441
基于改进天牛群优化的 DBSCAN 聚类算法	0.857	10.958
基于改进粒子群优化的 K-means 聚类算法	0.824	11.389

分析表 3 得出，分离系数用于度量云数据模糊聚类结果的优劣程度，其值无限趋近 1，聚类效果更突出；分离熵衡量聚类结果的紧凑性和分离性，其值越小，表示云数据模糊聚类效果越好。云计算模糊聚类算法在这两个指标上的表现优于对比算法，分离系数达到最大，其值为 0.891；分离熵最小，为 10.441。充分验证了云计算模糊聚类算法在云数据模糊聚类上的性能优势，显示出其在实际应用中的潜在价值和优越性。因为所提方法通过在搜索空间中不断探索和开发，找到更合适的聚类中心位置。优化后的聚类中心能够更好地代表各类数据的特征，使数据点与聚类中心之间的匹配更加准确，进而实现了不同类型云数据的准确区分，提高了模糊聚类性能。

在 3 组云数据集上进行模糊聚类分析后，通过聚类散点图可视化呈现不同类型云数据的分布情况，实验结果如图 3 所示。

轮廓系数结合了簇内紧密度和簇间分离度，其取值范围在－1 到 1 之间。值越接近 1，表示样本与其所属簇中的其他样本越相似，同时与其他簇中的样本越不相似，即聚类效果越好；值接近 0 表示样本可能位于两个簇的边界上；值接近－1 则表示样本可能被分配到了错误的簇。分析图 3 得出，对于组 3 数据集，经过计算，其平均轮廓系数为 0.78，这表明在该数据集上，云计算模糊聚类算法的聚类效果较好，样本在簇内的分布较为紧凑，且不同簇之间的分离度较高。相比之下，对比算法（如基于改进天牛群优化的 DBSCAN 聚类算法、基于改进粒子群优化的 K-means 聚类算法）在组 3 数据集上的平均轮廓系数为 0.65，明显低于云计算模糊

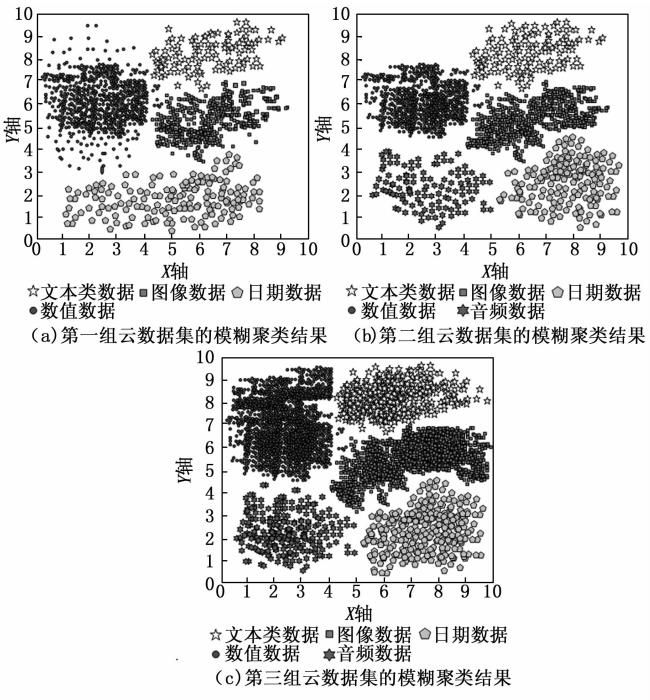

图 3　模糊 K-means 聚类散点图

聚类算法，进一步验证了云计算模糊聚类算法在云数据模糊聚类上的优势。这是因为所提方法在模糊 K-means 聚类算法引入了隶属度概念，能够灵活处理云数据样本在不同类别间的归属关系。算法根据数据点与聚类中心的距离以及隶属度来确定数据点所属的类别，使不同类别的数据在特征空间上形成了明显的区分。

选取加速比指标评估海量云数据模糊聚类并行化处理能力，加速比是衡量并行计算性能的重要指标，通过观察不同数据集下计算节点数量变化时加速比的变动情况，能够直观评估基于 MapReduce 的海量云数据模糊聚类并行化处理能力。了解随着计算节点的增加，系统处理云数据模糊聚类任务的速度提升程度，判断该并行化方法是否能够有效利用增加的计算资源来提高整体处理效率。在不同规模云数据集上，计算节点数与加速比的变化关系分析结果如图 4 所示。

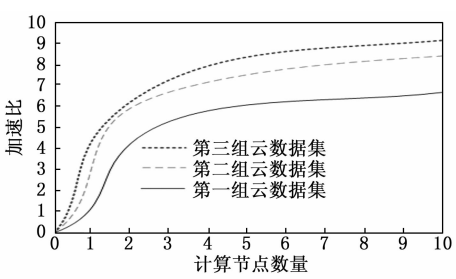

图 4　3 个数据集下计算节点数量与加速比关系分析

分析图 4 得出，随着计算节点数量的增加，加速

比指标呈现出持续上升的趋势，这说明增加计算节点可以有效提升云数据模糊聚类的并行处理效率。因为在基于 MapReduce 的海量云数据模糊聚类并行化处理中，Map 阶段将融合后的云数据进行切分并分配到不同从节点上，多个从节点可以同时进行云数据模糊聚类的计算任务。增加计算节点意味着更多的计算资源被投入到聚类任务中，能够并行处理更多的数据切片，从而有效提升云数据模糊聚类的并行处理效率，使加速比不断提高。

3 结束语

设计基于云计算技术的海量云数据模糊聚类算法，通过构建云计算数据分析框架，并结合随机森林算法、MapReduce 模型、模糊 K-means 算法以及量子粒子群算法，实现了对高维、复杂云数据的高效处理。随机森林算法在数据融合阶段有效整合了来自异构数据源的信息，提升了数据的全面性和准确性；MapReduce 模型通过数据切分和分布式计算，显著提高了数据处理的速度和效率；模糊 K-means 算法增强了聚类结果对数据不确定性和模糊性的适应能力，使得簇内数据更加紧凑，簇间区分度更高；量子粒子群算法优化了初始聚类中心，进一步降低了聚类误差，提高了聚类的精度和稳定性。未来，可进一步探索更先进的云计算技术和优化算法，以应对更大规模、更复杂云数据的挑战，推动云数据挖掘和分析技术的发展。

参考文献：

[1] 郑冬花，叶丽珠，隋 栋，等. 云计算环境中面向大数据的改进密度峰值聚类算法 [J]. 济南大学学报（自然科学版），2022，36 (5)：592－596.

[2] 田青云，文 成，徐 良. 基于云计算的数据挖掘聚类算法研究 [J]. 长江信息通信，2024，37 (9)：203－205.

[3] 文萍芳. 基于数据挖掘的电能表云端数据自适应聚类方法 [J]. 九江学院学报（自然科学版），2023，38 (1)：76－80.

[4] 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.

[5] SABERI H，SHARBATI R，FARZANEGAN B. A gradient ascent algorithm based on possibilistic fuzzy C-Means for clustering noisy data [J]. Expert Systems with Applications，2022，191 (Apr.)：1－20.

[6] 代少升，边志奇，袁中明. 结合软约束的演化数据流模糊聚类算法 [J]. 重庆邮电大学学报（自然科学版），2024，36 (2)：287－298.

[7] 张文宇，治 瑜，秦乐. 基于改进天牛群优化的 DBSCAN 聚类算法 [J]. 统计与决策，2022，38 (10)：20－25.

[8] 李艳霞，张艳芳. 基于改进 PSO 粒子群及 K-Means 聚类算法的物联网数据挖掘查全优化研究 [J]. 长江信息通信，2024，37 (1)：155－157.

[9] 张嘉旭，王 骏，张春香，等. 基于低秩约束的熵加权多视角模糊聚类算法 [J]. 自动化学报，2022，48 (7)：1760－1770.

[10] 张 丽，刘玉洁. 基于云计算的大数据分类挖掘算法研究 [J]. 信息与电脑（理论版），2023，35 (1)：72－74.

[11] 徐 敏，胡 聪，王 萍，等. 基于云计算技术的大规模数据存储策略研究 [J]. 微型电脑应用，2022，38 (4)：80－83.

[12] SONTI N，RUKMINI M S S，P. V R. Enhancing nano grid connectivity through the AI-based cloud computing platform and integrating recommender systems with deep learning architectures for link prediction [J]. Bulletin of the Polish Academy of Sciences：Technical Sciences，2024，72 (4)：150113.

[13] DAI J Q，XU W，OUYANG F F. Precision analysis of noncircular gears based on CNC machining technology under cloud computing platform [J]. Shock and Vibration，2022，2022 (1)：1359084.

[14] 丁嘉伟，冯乃勤. 云计算平台下基于小波分解的数据同步采集 [J]. 计算机仿真，2024，41 (1)：390－394.

[15] 蒋大锐，徐胜超. 基于改进 PSO-Means 算法的大数据聚类处理方法 [J]. 吉林大学学报（信息科学版），2024，42 (3)：430－437.

[16] 任建超，冯彦军，吕 扬，等. 基于 K-Means 模糊聚类的采场覆岩运移微震能量簇时空分析 [J]. 能源与环保，2023，45 (3)：308－312.

[17] LASEK P，RZSA W，ANNA KRóL. Aggregations of fuzzy equivalences in K-means algorithm [J]. Procedia Computer Science，2024，246：830－839.

[18] 吴辰文，王莎莎，曹雪同. 结合柯西分布和蚁狮算法改进的模糊聚类算法 [J]. 计算机工程与应用，2023，59 (17)：91－98.

[19] 李姜超，谢一航，李 辰，等. 基于改进 K-means 算法的科研仪器机时智能计算系统 [J]. 微型电脑应用，2024，40 (10)：156－160.

[20] REN Z，CHEN M. Hierarchical normal wiggly hesitant fuzzy K-means clustering algorithm [C] //International Conference on Computing，Control and Industrial Engineering. Singapore：Springer Nature Singapore，2024：515－525.

[21] 孙 瑾，宋娜娜，王 璐，等. 分布式的 KBB 索引树多关键词模糊排序搜索方案 [J]. 电子与信息学报，2025，47：1－11.