基于云计算模型的物联网边缘数据挖掘技术研究

梁亮
正德职业技术学院，江苏 南京 211106

摘要：本研究聚焦于基于云计算模型的物联网边缘数据挖掘技术，首先概述了数据挖掘技术的基本流程，包括数据集的选择、预处理、发掘和模式评估。接着，文中深入探讨了云计算的关键技术，涉及虚拟化技术、数据处理与编程模型构建技术，以及云计算典型平台的分析。研究重点放在聚类算法上，包括对 LDCK−K−means 算法的详细步骤分析和 ALDCK−means 算法的实验验证。最后，探讨了在 Hadoop 环境下 ALCDK−means 算法的设计与实践，希望本次研究能够为物联网边缘数据挖掘技术提供一个全新的视角和解决方案。

关键词：云计算模型；物联网边缘数据；聚类算法

Study on Edge Data Mining Technology Based on Cloud Computing Model

Liang,Liang

Zhengde Polytechnic, Nanjing, Jiangsu, 211106

Abstract: This study focuses on edge data mining technology based on the cloud computing model. Firstly, the basic process of data mining technology is outlined, including the selection, preprocessing, discovery, and pattern evaluation of datasets. Next, the paper delves into key technologies of cloud computing, involving virtualization technology, data processing, programming model construction technology, and analysis of typical cloud computing platforms. The research emphasizes clustering algorithms, providing a detailed analysis of the LDCK-K-means algorithm and experimental verification of the ALDCK-means algorithm. Finally, the design and implementation of the ALCDK-means algorithm in the Hadoop environment are discussed, aiming to offer a fresh perspective and solution for edge data mining technology in the Internet of Things.

Keywords: Cloud computing model; Edge data in the Internet of Things; Clustering algorithm

DOI: 10.62639/sspis01.20240101

物联网技术，作为当今时代的重要技术之一，通过网络平台实现了设备间的高度互联和智能化管理。这种技术的核心在于其能够实现信息的实时共享和处理，极大地增强了系统的响应能力和处理效率[1]。随着物联网技术的不断深入和完善，其在智能家居、工业自动化、城市管理等领域的应用越来越广泛，显著提升了生产效率和生活质量。与此同时，数据挖掘作为一种有效的信息处理手段，能够在海量的物联网数据中发掘潜在的价值和模式，为决策提供科学依据。通过精准的数据分析和智能算法，物联网结合数据挖掘技术，不仅能提高问题解决的效率和质量，还能预见和预防潜在的问题，为实现智能化管理和运维提供强有力的支撑。

一、数据挖掘技术应用的基本流程分析

数据挖掘技术应用的基本流程可分为四个关键阶段，分别为数据集的初步选择、数据集的预处理、数据发掘、以及模式评估，每个阶段对于数据挖掘过程都是必不可少的，他们共同确保了整个数据挖掘过程的有效性和准确性[2]。

（一）数据集的初步选择

这是数据挖掘的起始点，这一步包括对原始数据源的识别和选择，主要任务是确定数据挖掘的目标和需求，以及可用的数据资源。这包括从大型数据库、数据仓库或其他数据集中识别出最相关的数据集，以便进行进一步的分析。在这个阶段，重要的是要对数据的类型、规模和质量有一个基本的了解，这有助于后续的数据处理和分析。例如，一个零售企业可能会选择销售记录、客户反馈和市场趋势数据作为其数据挖掘项目的数据集。

（二）数据集的预处理

数据预处理是数据挖掘过程中至关重要的一步，它包括清洗、整合、变换和规约数据等多个子步骤。清洗数据涉及到处理缺失值、消除噪声和纠正不一致性，以确保数据质量。数据整合则是将多个数据源合并到一个一致的数据集中[3−5]。变换数据通常涉及到规范化和聚合，以使数据适用于挖掘任务。数据规约则是减少数据量但同时保持数据的完整性，以提高数据挖掘效率。这个阶段的目的是创造一个干净、一致和相关的数据集，为数据发掘提供良好的基础。

表 1 数据预处理的主要步骤及具体任务

步骤	任务描述
数据清洗	处理缺失值、消除噪声和纠正不一致性以确保数据质量。
数据整合	将多个数据源合并到一个一致的数据集中。
数据变换	包括规范化和聚合，使数据适用于挖掘任务。
数据规约	减少数据量，同时保持数据的完整性，提高数据挖掘效率。

（稿件编号：IS-24-1-1001）

（三）数据发掘

这个阶段是数据挖掘过程的核心，涉及到应用各种技术从预处理过的数据中提取有用信息和知识。常用的数据挖掘技术包括分类、回归、聚类、关联规则学习、异常检测和序列模式发现等。例如，在银行信用卡欺诈检测中，可以应用分类算法来识别可能的欺诈交易；在市场篮子分析中，关联规则学习用于发现顾客购买行为中的模式。在这个阶段，选择正确的挖掘算法和调整算法参数至关重要，因为它们直接影响到挖掘结果的质量和效果。

（四）模式评估

最后一个阶段是对挖掘出的模式和知识进行评估和解释。在这一阶段，分析师需要评估挖掘结果的有效性、可靠性和实用性。这涉及到使用统计方法、可视化技术和专家知识来评估和解释模式。此外，还需要考虑挖掘结果的业务相关性和可操作性[6-7]。例如，在一个零售企业的商品推荐系统中，分析师不仅需要评估推荐模型的准确性，还需要考虑其对业务目标的贡献。这个阶段的目标是确保挖掘出的知识对决策支持具有真正的价值。下图 1 为数据挖掘环节的具体工作开展流程

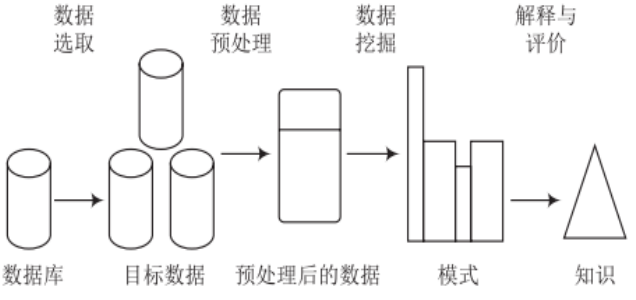

图 1 数据挖掘技术的应用流程

二、云计算关键技术分析

云计算作为信息技术领域的一项革命性进展，它所涉及的关键技术和应用已深入到了我们日常生活的各个方面。在云计算的众多关键技术中，虚拟化技术、数据处理与编程模型构建技术以及云计算典型平台的分析是最为核心和基础的三个方面。

（一）虚拟化技术

虚拟化技是云计算的基石，它通过抽象化的方式将物理资源转换为可灵活分配和管理的虚拟资源。虚拟化技术使得硬件资源的使用效率大幅提升，同时也为云计算提供了极高的灵活性和可扩展性。具体来说，它借助于创建虚拟机来模拟完整的硬件系统，包括虚拟的 CPU、内存、硬盘和网络接口等。这些虚拟机可以运行在单一或多个物理服务器上，而且彼此之间相互隔离，确保了安全性和稳定性。虚拟化技术的应用不仅限于服务器虚拟化，还包括存储虚拟化、网络虚拟化等，为云计算的实现提供了强大的支持。

表 2 虚拟化技术的关键特征和应用

特征 / 应用	描述
资源转换	将物理资源抽象化为虚拟资源，便于灵活分配和管理。
效率提升	提升硬件资源的使用效率，优化资源配置。
灵活性和可扩展性	通过虚拟化技术提高云计算的灵活性和可扩展性。
虚拟机	模拟完整硬件系统的虚拟环境，包括 CPU、内存、硬盘和网络接口等。
运行环境	虚拟机可以运行在单一或多个物理服务器上，彼此隔离保障安全稳定。
应用范围	包括服务器虚拟化、存储虚拟化、网络虚拟化等。
对云计算的支持	为云计算实现提供基础支撑和强大的技术支持。

（二）数据处理与编程模型构建技术

云计算环境中需要产生并存储海量的数据，那么到底应该如何有效地处理这些数据呢。为了应对这一挑战，MapReduce 编程模型这一类数据处理模型和编程框架也就出现了，这些模型能够对大规模数据集进行分布式处理，通过 Map（映射）和 Reduce（归约）两个步骤并行处理大量数据[8-9]。此外，为了适应不同类型的应用需求，还发展出了多种数据存储技术，比如关系型数据库、NoSQL 数据库和对象存储，这些技术应用于云计算平台中，既保证了数据处理的高效性，也为复杂的云应用提供了支持。

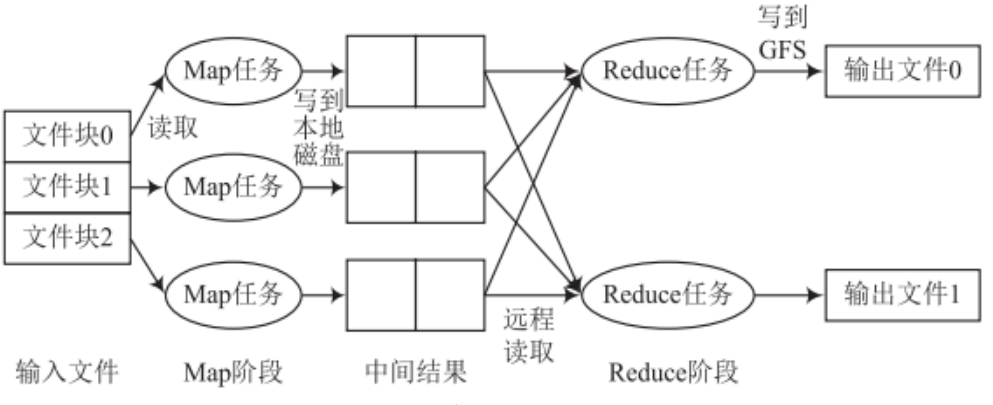

图 2 并行编程模型

（三）云计算典型平台

云计算平台通常分为三种服务模式：基础设施即服务（IaaS），平台即服务（PaaS）和软件即服务（SaaS），每种模式都有其特定的功能和应用场景[10]。例如，IaaS 主要用于提供虚拟化的计算资源，如 Amazon Web Services (AWS) 和阿里云; PaaS 则提供应用开发和部署的平台，如 Google App Engine；而 SaaS 可以直接向用户提供可用的软件应用，如 Salesforce。这些平台极大地简化了 IT 资源的管理，降低了企业的运营成本，同时也推动了新型服务和商业模式的创新。

三、聚类算法分析

（一）聚类算法的基本介绍

聚类算法是一种在无监督学习领域广泛使用的数据分析方法，其主要目标是将数据集中的对象分组，使得组内成员之间的相似性高于组间成员的相似性。在聚类分析中，数据被划分为若干个称为“簇”的子集，这些簇的构成基于数据特征的内在结构和相似性度量。聚类算法广泛应用于各个领域，包括但不限于市场细分、社交网络分析、计算生物学以及图像分割等。

首先，需要了解聚类的基本概念和数学表达。在聚类中，给定一个数据集 $D=\{x_1, x_2, \ldots, x_n\}$，其中每个 x_i 是一个具有多个特征的数据点，聚类的目标是将 D 划分为 k 个簇 $S=\{S_1, S_2, \ldots, S_k\}$，满足以下条件：每个簇 S_i 至少包含一个数据点，且 D 中的每个数据点属于且仅属于一个簇。聚类的质量通常通过内部凝聚度和外部隔离度来评估，内部凝聚度衡量簇内成员之间的紧密程度，而外部隔离度衡量不同簇之间的分离程度。

在聚类算法中，最常用的一种算法是 K- 均值聚类（K−means）。K- 均值聚类的目标是最小化每个点到其所属簇中心的距离之和，可以通过下面的优化问题来表达：$\min_S \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|$ 其中，μ_i 是簇 S_i 的中心，通常取为簇内所有点的均值。K- 均值算法通过迭代优化来解决这个问题，首先随机选择个初始中心，然后迭代进行两个步骤：(1) 将每个数据点分配给最近的簇中心；(2) 更新每个簇的中心为该簇内所有点的均值。这个过程重复进行，直到簇中心不再发生变化。

然而，K- 均值算法有一些局限性，例如它需要预先指定簇的数量，且对于非球形的簇效果不佳。为了克服这些局限，有许多其他的聚类算法被提出，例如，DBSCAN（Density−Based Spatial Clustering of Applications with Noise）是一种基于密度的聚类算法，它不需要预先指定簇的数量，并且能够识别任意形状的簇。DBSCAN 的基本思想是从某个核心点开始，如果这个点的 ϵ- 邻域内有足够多的点（即超过某个阈值 $MinPts$，则形成一个簇。算法继续递归地扩展这个簇，将 ϵ- 邻域内的所有密集点纳入簇中。这个过程不断重复，直到没有新的点可以加入任何簇。

在实际应用中，选择合适的聚类算法通常依赖于数据的特性和分析的目的，不同的聚类算法有各自的优缺点，因此在实际问题中需要根据具体情况选择最适合的方法。例如，对于具有噪声和不规则形状簇的复杂数据集，基于密度的聚类方法（如 DBSCAN）可能更合适。而对于大规模数据集，可能需要选择更高效的算法，如基于采样或近似的聚类方法。此外，评估聚类结果的有效性也是一个重要的研究方向，常见的评估指标包括轮廓系数（Silhouette Coefficient）、Davies−Bouldin Index 等。

（二）LDCK−K−means 算法的具体步骤分析

LDCK−K−means 算法在传统 K−means 算法基础上进行了改进，主要目的是提高数据聚类的准确性和效率。该算法的具体步骤如下：1）初始化：选择 K 个初始聚类中心。这些初始中心可以是随机选择的，或者使用一些启发式方法选择以提高算法的效率。2）分配数据点：将每个数据点分配给最近的聚类中心。这一步的目的是为了最小化每个数据点与其所属聚类中心之间的距离。3）更新聚类中心：计算每个聚类中所有数据点的均值，并将该均值设为新的聚类中心。4）迭代过程：重复步骤 2 和步骤 3 直到满足收敛条件。收敛条件可以是聚类中心的变化小于一个阈值，或者达到预设的迭代次数。5）聚类结果：输出最终的聚类中心和每个数据点的聚类标签。

LDCK−K−means 算法的关键在于如何选择初始聚类中心和如何计算数据点与聚类中心之间的距离。在实际应用中，可能还需要考虑数据的特殊性，比如数据的维度、分布特征等，从而对算法进行相应的调整或优化。此外，算法的性能也受到数据规模和计算资源的限制，因此在大规模数据集上应用时，可能需要采用分布式计算等技术以提高算法的可扩展性和效率。

（三）ALDCK−means 算法实验

ALDCK−means 算法是一种改进的聚类算法，用于数据挖掘和机器学习领域，该算法是对传统 K−means 算法的扩展，在此基础上引入了自适应学习和动态聚类键（ALDC）的概念。

实验开始前，首先需要准备和预处理数据集，具体包括数据清洗、标准化以及可能的维度降低步骤。若据集包含缺失值或异常值，需要对其进行处理或移除，确保后续分析的准确性，而数据标准化是为了避免不同量级的特征对聚类结果产生不平衡的影响，在预处理完成后，就可以开始实施 ALDCK−means 算法了。

ALDCK−means 算法的第一步是初始化，即选择初始聚类中心，涉及的步骤有随机选择、基于数据分布的启发式方法或其他先进的初始化策略。在初始化聚类中心之后，ALDCK−means 算法进入其核心迭代过程，在每次迭代中，算法首先根据当前的聚类中心对数据点进行分配，即基于数据点到聚类中心的距离，每个数据点被分配到最近的聚类中心，形成临时的聚类。

接下来是更新聚类中心，在传统的 K-means 算法中，这一步需要通过计算每个聚类内数据点的均值来完成。然而，在 ALDCK-means 算法中，可以引入自适应学习机制，此时聚类中心的更新不仅取决于聚类内的数据点，还可能受到数据点的密度、分布特征或者先验知识等其他因素的影响。

此外，ALDCK-means 算法中的“动态聚类键”允许算法在运行过程中调整聚类的数量，这是通过监测聚类的质量和特征来实现的。例如，如果两个聚类非常相似，它们可能会合并；如果一个聚类内部存在显著的子群体，它可能会分裂，这种动态调整使得 ALDCK-means 算法在处理复杂或不均匀分布的数据集时更加灵活和有效。

整个迭代过程持续进行，直到满足停止条件，这通常是聚类中心的变化低于某个阈值，或者达到了预定的迭代次数，完成迭代后，最终的聚类结果被输出，这包括每个数据点的聚类分配以及聚类中心的位置。

在实验的最后阶段，需要对 ALDCK-means 算法的性能进行评估，比较它与传统 K-means 算法或其他聚类方法的结果，评估聚类的质量，并分析算法的稳定性和可扩展性。此外，实验可能还包括敏感性分析，即探究不同的初始化策略或参数设置对算法性能的影响。

表 3 ALDCK-means 算法实验的关键步骤和内容

实验阶段	描述
数据预处理	包括数据清洗、标准化和可能的维度降低，处理缺失值或异常值，确保数据集的准确性和一致性。
初始化聚类中心	选择初始聚类中心，可以是随机选择、基于数据分布的启发式方法或其他先进的初始化策略。
核心迭代过程	每次迭代中，将数据点基于距离分配到最近的聚类中心，形成临时的聚类。
更新聚类中心	除了计算聚类内数据点的均值外，还引入自适应学习机制，考虑数据点的密度、分布特征等因素影响。
动态聚类键	在运行过程中根据聚类的质量和特征动态调整聚类数量，如合并相似聚类或分裂内部有显著子群体的聚类。
停止条件判断	判断迭代过程是否满足停止条件，如聚类中心变化低于阈值或达到预定迭代次数。
输出聚类结果	输出最终的聚类结果，包括每个数据点的聚类分配和聚类中心的位置。
性能评估	评估 ALDCK-means 算法与传统 K-means 或其他聚类方法的性能比较，包括聚类质量、稳定性和可扩展性分析。
敏感性分析	探究不同的初始化策略或参数设置对 ALDCK-means 算法性能的影响。

（四）Hadoop 背景下的 ALCDK-means 算法设计与实践

在 Hadoop 环境下，ALCDK-means 算法的优化使其更加高效地处理大规模数据集。在这种改进的计算方法中，每一个数据点的关键参数，如高密度最小距离、局部密度、核心距离等，都独立计算，保证了数据完整性和计算的精确性。基于这些优势，进一步将此算法并行化，使之能够有效地处理大量的数据信息。在这种并行化的框架下，每一个迭代步骤都对应着特定的数据分析任务，同时生成相关的密度最小距离值和局部密度值。每个计算阶段都呈现出独特的计算特性和步骤。经过并行化优化后的 ALCDK-means 算法执行流程具体如下：① 初始化集群处理，将聚类数据传输至分布式系统平台；② 基于平台读取聚类数据，并解析为适合处理的格式；③ 各个独立的区域性状态分别处理，计算每个数据点的高密度最小距离值和局部密度值；④ 分析所有数据点，筛除噪声数据，按降序排列，最后选定合适的聚类中心。

四、结语

综上所述，本研究通过综合分析数据挖掘的基本流程、云计算的关键技术以及聚类算法的应用，为基于云计算模型的物联网边缘数据挖掘技术提供了全面的理论基础和实践指导。通过对 LDCK-K-means 和 ALDCK-means 算法的深入研究，以及在 Hadoop 环境下的算法实践，本研究不仅展示了云计算在物联网数据处理中的应用潜力，而且为未来的研究和开发提供了新的思路和方法。随着物联网和云计算技术的不断进步，期待这一领域能够产生更多创新和突破，为智能数据处理和分析贡献更大力量。

参考文献：

[1] 宋文彬．基于云计算的数据挖掘平台架构及其关键技术研究 [J]．电子技术与软件工程，2021，(03)：211-212.

[2] 张启．基于大数据、云计算和物联网传感器技术的畜牧业信息化研究 [J]．农家参谋，2019，(18)：142.

[3] 陈琛．基于云计算物联网数据挖掘模式的构建 [J]．信息与电脑（理论版），2018，(23)：119-120.

[4] 程俊静．基于物联网云平台的海上船舶交通数据挖掘技术研究 [J]．舰船科学技术，2018，40 (18)：58-60.

[5] 肖露．基于大数据、云计算和物联网传感器技术的畜牧业信息化研究 [J]．农家参谋，2018，(05)：108.

[6] 刘建东．云计算下数据挖掘平台架构及其关键技术的探索 [J]．科技与创新，2017，(06)：128+132.

[7] 王国杰，于建涛，王立平，任建吉．地理时空大数据算法平台的研究 [J]．工业控制计算机，2023,36(06):113-114+117.

[8] 郑琳．基于物联网边缘计算的数据挖掘方法研究 [J]．无线互联科技，2022,19(15):140-142.

[9] 张永利，潘哲，李洋．基于数据挖掘的空中目标航迹特征提取技术研究 [J]．舰船电子对抗，2022,45(03):83-88.

[10] 向绍斌，涂水员．基于孤立森林算法与大数据挖掘的配电网故障距离估计方法 [J]．电气工程学报，2022,17(01):179-185.

（**作者简介**：梁亮（1981-02），男，汉族，籍贯 江苏淮安，硕士，正德职业技术学院实验师，研究方向：计算机应用，云计算与大数据应用，网络安全及应用，电子商务应用。
基金项目：江苏高校哲学社会科学研究项目，“课程思政”视域下高职院校专业课教师与辅导员协同育人体系研究，2022SJYB0908。）