大数据环境下数据挖掘算法优化研究

薛乔尉[1] 陈加博[1] 吕萌[2] 麻亮[3] 董珊珊[4] 付宇航[5]

1. 长春光华学院，中国·吉林 长春 130033

2. 吉林省白山市临江市苇沙河镇中心学校，中国·吉林 白山 134600

3. 吉林省白山市临江市蚂蚁河乡中心学校，中国·吉林 白山 131500

4. 长春市第一〇三中学校，中国·吉林 长春 136400

5. 长春新区启明学校，中国·吉林 长春 136500

摘要：数字技术推动大数据渗透多领域，其海量、高维、动态特征使传统数据挖掘算法面临效率低、精度波动、动态适配不足等问题。本文从数据预处理、并行计算融合、智能优化嵌入、动态适应构建四维度优化算法，经实验验证，优化后算法在效率、精度与动态适配性上显著提升，为大数据挖掘落地及行业数字化转型提供支撑。

关键词：大数据；数据挖掘算法；并行计算；动态适应机制；智能优化

Research on Optimization of Data Mining Algorithms in the Big Data Environment

Xue Qiaowei[1], Chen Jiabo[1], Lv Meng[2], Ma Liang[3], Dong Shanshan[4], Fu Yuhang[5]

1. Changchun Guanghua College, China Jilin Changchun 130033

2. Weishahe Town Central School, Linjiang City, Baishan City, Jilin Province, China Jilin Baishan 134600

3. Ant River Township Central School, Linjiang City, Baishan City, Jilin Province, China Jilin Baishan 131500

4. Changchun 103rd Middle School, China Jilin Changchun 136400

5. Qiming School, Changchun New Area, China Jilin Changchun 136500

Abstract: Digital technology promotes the penetration of big data into multiple fields, and its massive, high-dimensional, and dynamic features make traditional data mining algorithms face problems such as low efficiency, fluctuating accuracy, and insufficient dynamic adaptation. This article constructs a four-dimensional optimization algorithm from data preprocessing, parallel computing fusion, intelligent optimization embedding, and dynamic adaptation. Through experimental verification, the optimized algorithm significantly improves efficiency, accuracy, and dynamic adaptability, providing support for the implementation of big data mining and the digital transformation of the industry.

Keywords: Big data; Data mining algorithms; Parallel computing; Dynamic adaptation mechanism; Intelligent optimization

0 引言

数字经济时代，数据成为核心生产要素，物联网、云计算等技术催生 TB 甚至 PB 级数据，形成兼具海量、高速、多样特征的大数据环境。数据挖掘作为提取数据价值的关键技术，已在金融风控、电商推荐等领域发挥作用，但传统算法多适配小数据场景，处理大数据时易因计算开销剧增导致效率下降，也难以应对高维数据的“维度灾难”与动态数据的模型过时问题。这些适配性不足限制了数据价值释放，因此，结合大数据特征优化数据挖掘算法，成为突破技术瓶颈的关键，本文即围绕这一需求展开研究。

1 大数据环境下数据挖掘算法的应用特征与核心挑战

1.1 算法应用特征

大数据环境对数据挖掘算法的需求，本质上是“适配场景特征、平衡效率与精度”的需求，具体呈现为三个核心特征：一是可扩展性，大数据的海量特征要求算法能够处理 TB 甚至 PB 级数据，不能因数据规模增长导致性能急剧下降，需在数据量扩大时保持相对稳定的计算效率；二是时效性，高速产生的动态数据要求算法具备快速处理能力，能够在短时间内完成数据分析并输出结果，以满足实时决策需求；三是多源适配性，多样的数据源与数据类型

要求算法能够处理结构化、半结构化与非结构化数据的融合挖掘，避免因数据类型差异导致的信息丢失，同时在低价值密度数据中精准提取有用信息，提升挖掘结果的实用性[1]。

1.2 核心适配挑战

基于上述应用特征，传统数据挖掘算法在大数据环境下主要面临三方面挑战：首先是高维数据导致的计算与精度矛盾，大数据环境下的特征维度常达数千甚至数万维，传统算法在处理高维数据时，不仅计算量会随维度增加呈指数级增长，还易因冗余特征干扰导致挖掘精度下降，如何在降低维度的同时保留关键信息，成为算法适配的首要挑战；其次是动态数据导致的模型时效性不足，大数据环境中的数据常处于动态变化中，传统算法构建的静态模型难以实时更新，随着数据分布变化，模型精度会逐渐降低，如何让算法实时感知数据变化并动态调整模型，是保障挖掘效果的重要难题；最后是海量数据导致的计算效率瓶颈，传统算法多基于单机架构设计，内存与计算能力有限，在处理海量数据时易出现“内存溢出”或“计算超时”问题，即使部分算法可通过分批处理完成计算，也会因数据分割不合理导致精度损失，如何在分布式架构下实现高效计算与精度保障的协同，是算法落地的关键障碍[2]。

2 大数据环境下数据挖掘算法的优化方向与关键技术

针对大数据环境下数据挖掘算法的适配挑战，结合算法应用特征，本文从四个核心方向展开优化研究，通过技术创新实现算法与大数据场景的深度适配。

2.1 基于数据预处理的算法轻量化优化

数据预处理是降低算法计算开销、提升挖掘精度的基础环节，其核心思路是通过“减冗余、提质量”实现算法轻量化，主要包括特征选择与数据采样两项关键技术。在特征选择方面，针对高维数据的冗余问题，引入基于信息增益的 Relief-F 算法与基于正则化的 LASSO 算法：Relief-F 算法通过计算特征与类别标签的相关性，筛选出对分类结果贡献度高的特征，剔除无关与冗余特征；LASSO 算法则通过添加 L1 正则项，将部分特征系数压缩至零，实现特征的自动选择与降维。这两种算法的结合，既能避免单一方法的局限性，又能在降维过程中保留关键信息，减少算法的计算量。

在数据采样方面，针对海量数据的处理效率问题，采用基于分层抽样的改进方法：首先根据数据的类别分布或特征属性，将原始数据划分为多个层次；然后在每个层次内按照一定比例进行抽样，确保抽样数据的分布与原始数据一致，避免因抽样偏差导致的精度损失。同时，针对不平衡数据，引入 SMOTE 算法（合成少数类过采样技术），通过合成少数类样本，平衡数据分布，提升算法对少数类样本的识别能力。数据预处理技术的应用，使得算法在处理大数据时，既能减少计算开销（如降维后特征数量减少 50% 以上），又能保障挖掘精度（如分类算法的准确率提升 8%-12%），为算法的高效运行奠定基础。

2.2 基于并行计算框架的算法效率提升

并行计算框架是解决海量数据计算效率瓶颈的核心技术，其核心思路是将数据挖掘任务拆解为多个子任务，通过分布式节点的并行执行，提升整体计算效率。目前主流的并行计算框架包括 Hadoop 与 Spark，其中 Spark 因基于内存计算的特性，在实时数据处理方面表现更优，成为大数据挖掘算法优化的重要载体。在算法与 Spark 框架的融合优化中，针对不同类型的挖掘算法采用差异化的并行化策略：对于分类算法（如决策树、支持向量机），将数据按照特征维度或样本数量拆分为多个分区，每个分区分配至不同的计算节点，节点独立完成局部模型的训练，再通过参数聚合实现全局模型的构建；对于聚类算法（如 K-Means），采用“数据本地化”策略，将聚类中心与相近的样本数据分配至同一节点，减少节点间的数据传输量，同时通过迭代计算优化聚类中心，提升算法的收敛速度；对于关联规则挖掘算法（如 Apriori），将频繁项集的挖掘任务拆解为多个子任务，每个节点负责挖掘局部频繁项集，再通过频繁项集的合并与验证，实现全局频繁项集的挖掘。

并行计算框架的应用，显著提升了算法的计算效率：以 K-Means 算法为例，在处理 1000 万条样本数据时，传统单机版算法需要 6-8 小时完成计算，而基于 Spark 框架的并行化算法仅需 40-60 分钟，计算效率提升 8-10 倍；同时，通过合理的任务拆解与节点分配，算法在处理数据量增长时，仍能保持相对稳定的效率（如数据量增加 10 倍，计算时间仅增加 2-3 倍），有效解决了海量数据的计算效率瓶颈。

2.3 基于智能优化策略的算法精度协同

智能优化策略是平衡算法效率与精度的重要手段，其核心思路是通过嵌入启发式优化算法，对数据挖掘算法的参数或结构进行优化，提升算法的挖掘精度。目前常用的智能优化算法包括遗传算法、粒子群优化算法、蚁群算法等，这些算法具有全局搜索能力强、鲁棒性好的特点，能够有效解决传统算法易陷入局部最优的问题[3]。

在算法优化中，智能优化策略的应用主要体现在两个方面：一是参数优化，针对数据挖掘算法的关键参数（如K-Means的聚类数K、支持向量机的核函数参数），通过智能优化算法进行全局搜索，找到最优参数组合。例如，在支持向量机的参数优化中，采用粒子群优化算法，将参数（惩罚因子C、核函数参数σ）作为粒子的位置向量，以算法的分类准确率作为适应度函数，通过粒子的迭代更新，找到最优参数组合，避免传统网格搜索方法效率低、易陷入局部最优的问题；二是结构优化，针对决策树算法的结构（如树的深度、节点分裂阈值），采用遗传算法进行优化，将决策树的结构编码为染色体，以算法的预测误差作为适应度函数，通过选择、交叉、变异操作，优化决策树的结构，减少过拟合现象的发生。

智能优化策略的嵌入，有效提升了数据挖掘算法的精度：以决策树算法为例，未优化的算法在复杂数据集上的预测误差约为15%-18%，而通过遗传算法优化后的决策树，预测误差降低至8%-10%；同时，智能优化算法的全局搜索能力，使得数据挖掘算法在处理高维、异构数据时，仍能保持较高的稳定性，避免因数据特征变化导致的精度波动。

2.4 基于动态适应机制的算法场景适配

动态适应机制是解决算法对动态数据时效性不足的关键技术，其核心思路是通过实时感知数据分布变化，动态调整算法模型或参数，确保算法在动态场景下的挖掘效果。动态适应机制的实现主要包括两个环节：一是数据分布监测，通过滑动窗口技术或指数加权移动平均方法，实时采集最新数据样本，分析数据的均值、方差、类别分布等统计特征，判断数据分布是否发生显著变化（如变化幅度超过预设阈值）；二是模型动态更新，当监测到数据分布变化时，采用增量学习或在线学习方法更新模型：增量学习通过将新数据样本融入原有模型，避免重新训练模型导致的效率损失（如增量SVM算法仅需更新支持向量，无需重新计算所有样本）；在线学习则通过实时接收数据样本，逐次更新模型参数，确保模型与数据分布保持同步（如在线梯度下降算法，通过每次迭代更新参数，适应数据的动态变化）。

动态适应机制的应用，显著提升了算法对动态场景的适配性：以电商推荐系统中的协同过滤算法为例，未引入动态适应机制时，随着用户偏好变化，推荐准确率会在1个月内下降15%-20%，而引入动态适应机制后，通过实时监测用户行为数据变化并更新模型，推荐准确率的下降幅度控制在5%-8%以内；同时，动态适应机制的轻量化设计（如增量学习仅需处理新数据样本，计算量仅为重新训练的10%-20%），确保了算法在动态场景下的高效运行，避免因模型更新导致的效率损失。

3 优化算法的效果验证与实践价值

3.1 效果验证

为验证优化后数据挖掘算法的性能，选取金融风控、电商推荐、智慧城市三个典型应用场景，采用实际业务数据进行对比实验（实验环境为Spark分布式集群，包含5个计算节点，每个节点配置16GB内存、8核CPU）。实验以传统算法为对照组，优化算法为实验组，从计算效率、挖掘精度、动态适配性三个维度进行评估。

在计算效率方面，金融风控场景处理1000万条交易数据时，传统决策树算法的训练时间为120分钟，优化后的决策树算法（结合并行计算与数据预处理）训练时间缩短至35分钟，效率提升70.8%；电商推荐场景处理5000万条用户行为数据时，传统协同过滤算法的推荐响应时间为1.5秒，优化后的协同过滤算法（结合并行计算与动态适应机制）响应时间缩短至0.3秒，效率提升80%。

在挖掘精度方面，金融风控场景中，传统支持向量机算法的异常交易识别准确率为78.5%，优化后的支持向量机算法（结合智能优化与数据预处理）识别准确率提升至89.2%，精度提升13.6%；智慧城市场景中，传统K-Means算法对交通流量数据的聚类准确率为72.3%，优化后的K-Means算法（结合智能优化与并行计算）聚类准确率提升至85.6%，精度提升18.4%。

在动态适配性方面，电商推荐场景中，模拟用户偏好每月变化20%的动态场景，传统协同过滤算法的推荐准确率1个月后下降至65.3%，优化后的协同过滤算法（结合动态适应机制）推荐准确率1个月后仍保持在82.7%，动态适配性显著提升。

实验结果表明，优化后的数据挖掘算法在计算效率、挖掘精度、动态适配性三个维度均优于传统算法，能够有效适配大数据环境的特征需求。

3.2 实践价值

优化后的大数据挖掘算法，在实际应用中具有三方面核心实践价值：一是提升业务运行效率，通过并行计算与数据预处理技术，算法能够快速处理海量数据，减少业务等待时间，如金融机构采用优化后的算法，可将每日交易风控分析时间从4小时缩短至1小时，为风控决策提供更及时的支持；二是保障决策准确性，通过智能优化与动态

适应机制，算法能够在复杂、动态的大数据环境中保持较高的挖掘精度，如电商平台采用优化后的推荐算法，可将用户点击率提升 30% 以上，提升用户体验与平台收益；三是降低技术应用成本，优化后的算法在分布式架构下运行，无需依赖高端单机硬件，可通过普通服务器集群实现高效计算，同时动态适应机制减少了模型重新训练的频率，降低了计算资源消耗，为中小企业应用大数据挖掘技术提供了低成本路径。

4 结语

本文针对大数据环境下数据挖掘算法的适配挑战，从数据预处理、并行计算、智能优化、动态适应四方面探索优化路径，通过实验验证，优化后算法在计算效率、挖掘精度与动态适配性上均优于传统算法，可有效支撑金融、电商等领域的大数据应用。未来研究可进一步融合人工智能大模型提升非结构化数据处理能力，结合隐私计算保障数据安全，同时针对不同行业需求开发个性化方案，持续推动大数据挖掘技术赋能数字经济发展。

参考文献：

[1] 许丽娟，叶仕通. 非显著特征数据挖掘中 SOM 聚类算法的优化[J]. 计算机仿真，2023,40(09):497-501.

[2] 何庆，钟维坚，覃志智，林锋，唐苏东. 基于云计算的大数据聚类挖掘算法研究[J]. 中国新通信，2023,25(24):19-21.

[3] 吴玉凤. 大数据平台中基于深度学习的数据挖掘算法优化与系统设计[J]. 信息与电脑（理论版），2024,36(01):97-99.

基金项目：2025 年度长春光华学院“励新”计划项目课题名称：基于深度学习的交易行为特征分析与研究，课题编号：LXJH2025033。

作者简介：薛乔尉（1997.06-），男，汉族，吉林长春人，硕士研究生，助理研究员，研究方向：计算机科学技术。