基于 LOGO 计算框架开发的新型分布式机器学习算法库

梁展雄[1]，孙旭东[1]，蔡湧达[1]，张育铭[1]，麦朗杰[1]，何玉林[3]，黄哲学[1, 2, 3]
[1]（深圳大学，深圳 518060）
[2]（大数据国家工程实验室，深圳 518060）
[3]（人工智能与数字经济广东省实验室（深圳），深圳 518060）

摘要：LOGO 是一种新的分布式计算框架，与流行的 MapReduce 计算框架不同，LOGO 框架下的大数据分布式计算由两步操作完成，即 LO 操作在节点虚拟机内运行串行算法完成一个随机样本块的独立计算，产生局部计算结果；GO 操作将所有局部结果上传到主节点，在主节点内对局部结果做集成，得到大数据的近似计算结果。LOGO 计算框架执行迭代算法时，消除了节点间的数据通信，极大地提高了分布式计算的效率，降低了内存需求，提高了数据扩展性。本文介绍基于 LOGO 计算框架自主研发的一种新型分布式机器学习算法库 RSP-LOGOML。新型分布式计算由 LO 操作执行的串行算法和 GO 操作执行的集成算法两部分组成，LO 操作直接执行已有的机器学习串行算法，不需按 MapReduce 编程模型对算法重写，GO 操作对串行计算结果进行集成。本文阐述 LOGO 分布式计算的原理、算法库架构、串行算法封装和 GO 操作集成策略，展示 Spark 实现、App 应用开发和多种算法测试结果。

关键词：大数据分布式计算；分布式机器学习算法库；近似计算；Non-MapReduce 计算
中图分类号： 文献标志码 A doi: 10.12146/j.issn.2095-3135.20240224001

A New Distributed Machine Learning Library Developed Based on LOGO Computing Framework

LIANG Zhanxiong[1], SUN Xudong[1], CAI Yonda[1], ZHANG Yuming[1], MAI Langjie[1], HE Yulin[2], HUANG Zhexue[1,2,3],

[1] (Shenzhen University, Shenzhen 518060, China)
[2](National Engineering Laboratory for Big Data System Computing Technology, Shenzhen 518060, China)
[3](Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen 518060, China)

Corresponding Author: zx.huang@szu.edu.cn.

Abstract: LOGO is a new distributed computing framework using a Non-MapReduce computing paradigm. Under the LOGO framework, big data distributed computing is completed in two steps. The LO operation runs a serial algorithm in a number of nodes or virtual machines

来稿日期：2024-02-24 修回日期：2024-05-09
基金项目：深圳市基础研究重点项目(JCYJ20220818100205012)；广东省自然科学基金面上项目(2023A1515011667)；深圳市基础研究面上项目(JCYJ20210324093609026)
作者简介：梁展雄，硕士研究生，研究方向为大数据近似计算；孙旭东，博士研究生，研究方向为数据挖掘、近似计算、分布式计算和自然语言处理；蔡湧达，博士研究生，研究方向为机器学习和数据挖掘；张育铭，硕士研究生，研究方向为数据挖掘和大数据系统计算技术；麦朗杰，硕士研究生，研究方向为大数据近似计算；何玉林，博士，研究员，研究方向为大数据系统计算技术和多样本统计分析方法。黄哲学，博士，教授，研究方向为大数据系统计算技术以及面向大数据的数据挖掘和机器学习算法及其应用，E-mail: zx.huang@szu.edu.cn。

to process independently the random sample data blocks, generating local results. The GO operation uploads all local results to the master node and integrate them to obtain the approximate result of the big data set. The LOGO computing framework eliminates data communication between nodes during iterations of the algorithm, greatly improving computing efficiency, reducing memory requirements, and enhancing data scalability. This article proposes a new distributed machine learning algorithm library RSP-LOGOML under the LOGO computing framework. A new distributed computing is divided into two parts: the serial algorithm executed by the LO operation and the ensemble algorithm executed in the GO operation. The LO operation can directly execute existing serial machine learning algorithms without the need to rewrite them according to MapReduce. The GO operation executes ensemble algorithms of different kinds depending on the ensemble tasks. In this article, the principle of LOGO distributed computing is introduced first, followed by the algorithm library structure, the method for packaging existing serial algorithms and the ensemble strategy. Finally, implementation in Spark, App development, and the results of performance tests for various algorithms are demonstrated.

Key words: Big data distributed computing; Distributed machine learning library; Approximate computing; Non-MapReduce computing

Funding : Key Basic Research Foundation of Shenzhen (JCYJ20220818100205012), Natural Science Foundation of Guangdong Province (2023A1515011667), Basic Research Foundations of Shenzhen (JCYJ20210324093609026)

1 引 言

基于 MapReduce 计算框架的大数据智能分析系统，如：Hadoop MapReduce 和 Spark[1], 运行复杂迭代算法进行大数据分析时存在算法执行效率低、数据扩展能力差和内存瓶颈等卡脖子问题，原因是 MapReduce 计算框架在执行迭代算法时会产生大量的数据通信开销。另一个严重问题是需要对串行算法按 MapReduce 编程模型进行重写，复杂算法的编程难度大，运行效率低。因此，现有的分布式大数据计算系统普遍存在分析算法匮乏的现象，极大地制约了机器学习方法在大数据分析中的应用。

LOGO 是最近提出的一种新的 Non-MapReduce 分布式计算框架[2][3]，在随机样本划分（Random Sample Partition, RSP）大数据表达模型的支持下[4][5]，采用 LO 和 GO 两步操作完成对大数据的分布式计算。LO 操作在节点的虚拟机中运行一个串行算法对一个 RSP 样本数据块进行分析，得到大数据的一个局部结果。在这一步，LO 操作可以同时在许多虚拟机中执行同一种算法。对不同的 RSP 数据块做独立并行计算，得到多个局部结果。GO 操作将局部结果读取到主节点，在主节点运行集成算法对局部结果做融合操作，得到大数据的集成结果，作为大数据的近似计算结果。

与 MapReduce 相比，LOGO 计算框架在执行迭代算法做分布式计算时有如下优点：（1）算法执行效率高；（2）大数据计算不再受内存限制；（3）数据扩展能力强; (4) 迭代算法可以在 LO 操作中直接执行，不需按 MapReduce 编程模型重写。这些优点克服了 MapReduce 分布式计算的不足，提高了分布式大数据计算的效率和算法的多样性，丰富了分布式大数据分析系统的功能。

本文介绍一个新的基于 LOGO 计算框架开发的分布式机器学习算法库—RSP-LOGOML。该算法库在开源软件 Spark 上开发，机器学习串行算法采用开源库 Smile[1] 中的算法。为了便于算法库的使

[1] https://haifengl.github.io/

用、管理、维护和扩展，我们设计开发了算法库的组织架构、算子封装标准、算子的输入和输出数据结构以及 LO 操作和 GO 操作在应用程序（APPs）中的调用格式。这种标准化的算法库设计和开发极大地简化了大数据分析应用的编程复杂度。

我们选择了 7 个数据集对 RSP-LOGOML 算法库中的一些算法做了性能测试，并与 Spark 系统中的相同算法的性能做了对比。结果表明，RSP-LOGOML 算法库的算法运算效率远高于 Spark 算法的运算效率。有监督学习算法的精度率高于 Spark 同类算法的精度，原因是我们采用的是集成学习策略，Spark 算法产生是单个模型的结果。

本文余下的部分安排如下。第 2 节介绍了本文的一些相关工作，第 3 节介绍 RSP-LOGOML。实验部分见第 4 节。第 5 节讨论并分析了实验结果。第 6 节是本文的结论。

2 相关工作

分布式大数据计算平台是大数据存储与分析的重要支撑。以谷歌提出的 MapReduce 计算框架和编程模型为核心技术的主流大数据平台在大数据应用中发挥着重要作用。但是，随着数据集的增大和分析算法复杂度的增加，MapReduce 计算框架在执行迭代算法做分布式大数据分析时遇到了前所未有的挑战，具体挑战和原因在参考文献[2]中有详细阐述，本处不再赘述。为解决这个问题，学者最近提出了 Non-MapReduce 计算框架，消除了算法迭代过程中节点间的数据通信，提高了计算效率和数据扩展能力。本节对 Non-MapReduce 计算的相关技术作以概述，同时对机器学习算法库和集成学习的集成策略作简单回顾。

2.1 RSP 数据表达模型

Non-MapReduce 计算建立在 RSP 数据表达模型之上[4][5][6]，它将大数据文件表示成互不相交的随机样本块，称为 RSP 数据块。设 D 是一个大数据集文件，$D_1, D_2, \cdots, D_k$ 是 D 的 K 个 RSP 块，满足以下数学条件称 D_i 是 D 的随机样本:

（1）$D = \bigcup_{i=1}^{K} D_i.$

（2）$D_i \cap D_j = \varnothing, 1 \le i, j \le K, i \ne j.$

（3）$E[F_{D_i}(x)] = F_D(x), 1 \le i \le K,$

这里，$F_{D_i}(x)$ 和 $F_D(x)$ 分别表示 D_i 和 D 的累积分布函数。

可以看出，每个 RSP 数据块是 D 的一个随机样本，可以独立计算来估计 D 的统计特征。为了提高估计值的精度，随机抽取多个 RSP 数据块，独立计算多个估计值，根据中心极限定理，计算统计量的统计值作为 D 的估计值。从而奠定了 Non-MapReduce 计算的统计基础。为了解决了分布式大数据文件向 RSP 大数据表达模型的高效转换问题，Wei 等提出了两阶段的 RSP 块生成算法[7]。

2.2 LOGO 计算框架

LOGO 计算框架由深圳大学大数据技术与应用研究所开发[2][8]，是一种基于 RSP 数据表达模型的新型 Non-MapReduce 分布式计算框架。LOGO 计算框架将大数据分布式计算分为两个阶段：LO 阶段（局部阶段）和 GO 阶段（全局阶段），在 LO 阶段，串行的数据分析算法在每个 RSP 分区中独立地、并行地执行，该阶段不产生全局的数据交换。各个分区得到各自的局部结果。LO 阶段完成后，这些结果通过网络传输给到负责 GO 操作的节点上，在 GO 阶段中，该节点执行集成计算过程，得到全局结果。

整个过程中，数据通信在 LO 阶段和到 GO 阶段之间只发生一次。在迭代任务中，迭代过程在各个 RSP 分区独立执行，不进行全局的迭代。相对于 MapReduce 计算框架，大大降低了多次迭代过程产生的数据通信开销，因此 LOGO 框架更适合迭代任务。

LOGO 框架面向批处理的计算场景，考虑到 Spark 计算引擎优越的批处理计算性能，因此 LOGO 框架在 Spark 上实现。尽管 Flink 同样支持批处理计算[9]，相比于

Spark 具有一定的性能优势[10]，但其性能的优势更多地体现在对流数据的处理能力上[10][11]。

2.3 机器学习算法库

机器学习算法库是数据分析和挖掘的重要工具，流行的开源的机器学习算法库有 Weka、Scikit-learn 等，这些算法库包含的都是串行算法，在单机系统上运行，尽管算法丰富，但受到单机内存的限制，不能用于大数据分析。

主流的分布式大数据分析系统 Hadoop MapReduce 和 Spark 有自身的分布式机器学习库 Mahout[12]和 MLlib[13][14]。此外，Boden C 等使用 Spark 和 Flink 中的算子重新编写了一些常见的机器学习算法[15]，这些分布式算法提高了计算效率和数据扩展性。Smart-MLlib[16]采用类似于 MapReduce 的 Smart 框架实现了常用的机器学习算法，在计算时减少了中间数据的生成，并使用共享内存降低了数据通信的开销。Dislib[17]提供编程注解功能，用户使用这些注解对计算任务进行标记，对任务实现并行化计算，无需进行并行化编程。计算过程中数据为划分子块，并映射为多个子任务并行计算，再将子任务的结果合并得到最后的结果。

上述分布式机器学习算法库的算法在 MapReduce 或者改进的 MapReduce 计算框架上运行，计算效率低，数据扩展能力差。由于 MapReduce 编程的限制，可用算法有限，不能满足大数据分析多样性的需求。

2.4 集成学习

LOGO 计算框架采用的是集成学习策略，即：LO 操作执行一个机器学习算法对一组 RSP 数据块做并行计算，生成一组子模型；GO 操作执行一个集成算子对子模型做集成计算，生成集成结果。

常用的集成学习方法 Bagging，Boosting，Stacking[18][19][20][21][22][23]和新的集成方法 Dagging、Random Subspace[19]采用不同的集成策略计算集成结果。分类算法集成常用投票法，例如：多数投票、加权投票和 Approval 投票法[22]等。回归算法集成常用平均法，例如：加权平均、简单平均等[24]。聚类算法集成的传统方法有协关联矩阵法、超图划分法和基于二次互信息法等[25][26][27]，但是这些方法无法扩展到大数据聚类集成领域中。Mohammad 等最近提出的多随机样本分布式聚类集成方法采用了不同的集成策略，如：ACEM-GM、MSFC-GS 和簇球模型等[25][26][28]进行大数据聚类，为分布式聚类集成提供了新的解决方案。针对基于 MapReduce 的分布式频繁项集挖掘算法[12][29]计算效率低和数据扩展性差的问题，Sun 等提出的分布式集成方法[8]，减少通信开销获得近似的频繁项集挖掘结果。

3 算法库的设计与实现

本节介绍 RSP-LOGOML 算法库的设计和实现步骤，系统实现平台采用开源系统 Spark。在 Master-Slave share nothing 系统架构下，LOGO 计算框架如图 1 所示。首先，分布式大数据以 RSP 数据块形式随机存储在 Slave 节点上。对大数据的分析操作用两步完成，第一步是通过 LO 操作在 Slave 节点上对 RSP 数据块做计算，产生局部计算结果；第二步将 Slave 的局部结果上传到 Master 节点，用 GO 操作对局部结果做集成计算，产生最终的集成结果。

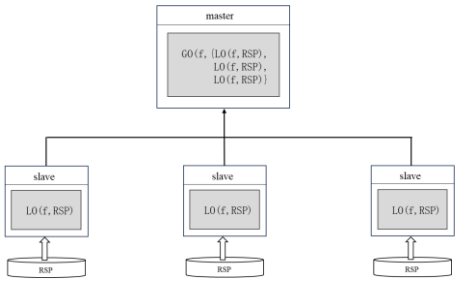

图 1 基于 Master-Slave 架构的 LOGO 计算框架

Fig. 1 LOGO computing framework based on Master-Slave share nothing architecture

基于 LOGO 计算框架的 RSP-LOGOML 算法库系统由算法库、算子库和 Apps 编程三部分功能组成：

1) 算法库部分包含所有机器学习算法的源码和可执行代码，源代码用 Scala 和 Java 编写，编译成独立的 Jar 包。

2) 算子库包含所有算法通过标准化封装的算子，在 Apps 中通过 LO 和 GO 操作调用执行。标准化部分包括算子名称、输入和输出数据结构、参数名及其排列顺序等，方便 Apps 开发。

3) Apps 编程分成三个主要部分，输入数据处理、LO 操作算子选择和 GO 操作集成算子选择。对于特殊的集成操作，可以在 Apps 中编写集成函数作为集成算子。

RSP-LOGOML 算法库系统架构如图 2 所示。

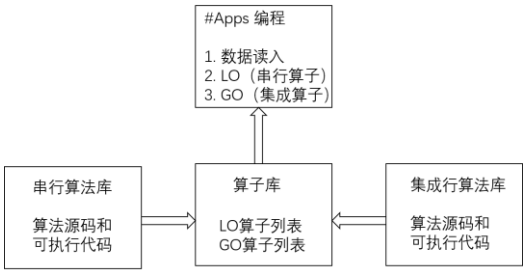

图 2　RSP-LOGOML 算法库架构

Fig. 2　RSP-LOGOML architecture

3.1　算法库

RSP-LOGOML 算法库可以封装用 Java 或 Scala 编写的任何数据分析算法，因为 Spark 上只能执行这两种语言的程序。我们首先选取了开源系统 Smile 中的机器学习算法作为首批机器学习算子。根据分析功能划分成分类、回归、聚类、关联规则、特征工程和集成策略等多类算法，如图 3 所示。算法源码在串行算法库中按类别分目录管理，方便代码维护和管理。

RSP-LOGOML 可以灵活增加和删除算法，新算法可以添加到现有类别中，也可以另增新的类别，如自然语言处理算法类别。算法库中的算法代码可以在外部独立开发和测试，确认无误后再加入算法库，降低算法的开发和测试难度。

3.2.　算子封装

LO 操作是应用程序调用串行算法的通用算子。用户通过 LO 操作调用串行算法发的算子并设置合理的参数，使用外部算法。

由于外部算法的源代码不是按统一的命名标准和数据结构开发，需要对算法做标准化封装，做成标准化算子，供 LO 操作调用，可以简化 Apps 的开发和调试难度，避免产生执行错误。

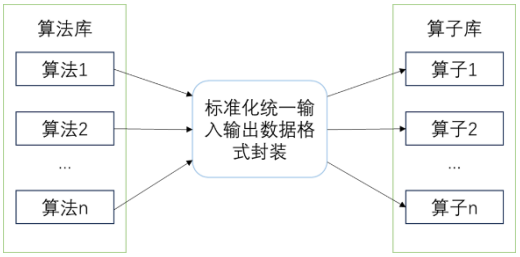

图4　算法封装示意图

Fig. 4　Diagram of algorithm encapsulation

如图 4 所示，算法的封装采用固定的标准化模式。首先设计新的算子类,每个算子类包含输入数据结构相同的算子。例如：分类的输入数据包括独立特征和类标特征。外部的分类代码不一定采用相同的输入数据格式。但是，系统的 RSP 数据块

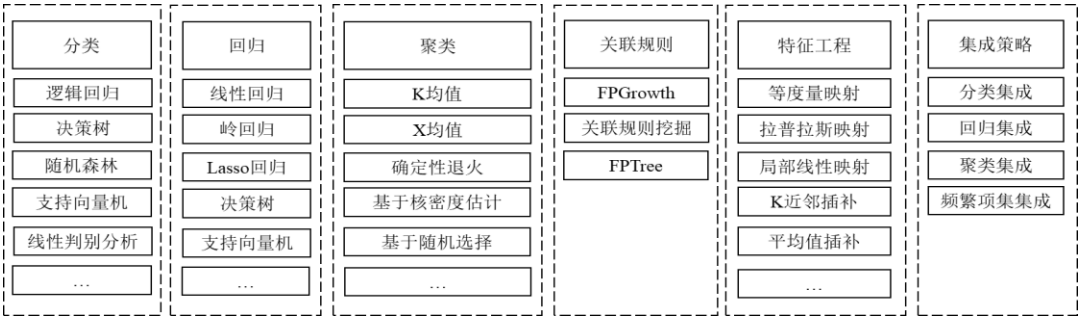

图3　算法库的算法分类

Fig. 3　Classification of algorithms in RSP-LOGOML

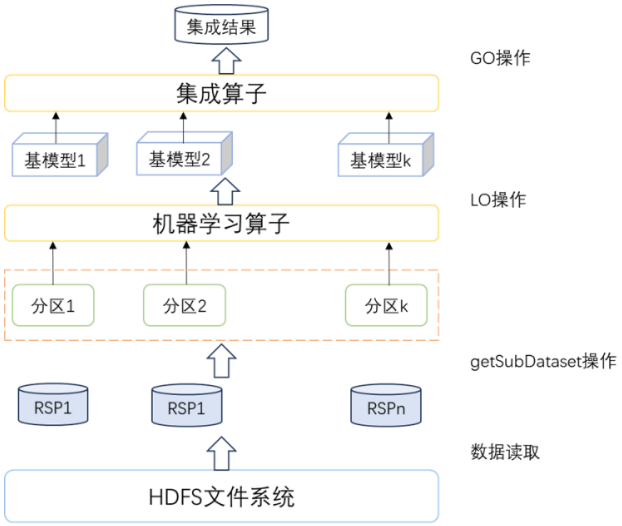

图 5　Apps 流程示意图

Fig. 5　Diagram of Apps procedure

格式是标准的，如果某个算法的输入数据格式与 RSP 数据块的格式不同，在算子的封装代码中要将 RSP 的数据块格式转换成算法源代码的输入数据格式，调用源代码进行运算。同理，如果源代码的输出数据格式与系统定义的输出数据格式不同，在封装代码中同样要完成数据格式转换。

LOGO 计算框架根据 RSP 数据模型的特点，设计了继承类 RspRDD 作为分布式数据的表达格式。RspRDD 也是 LO 操作的输入数据结构。同样，LO 操作的输出也用 RspRDD。根据算法的输出不同，RspRDD 的分区表达也不同。分区表达的结构由 GO 操作执行的集成算子自动识别，完成集成计算。

3.3　Apps 编程

算子库封装好后，用户可以用算子库中的算子编写应用程序 Apps，对分布式大数据做分析。

在 LOGO 计算框架下，Apps 的开发主要由三部分组成：数据预处理、LO 操作调用串行算子和 GO 操作调用集成算子 Apps 架构如图 5 所示。

本小节基于一个简单应用程序的例子来介绍 LO 操作和 GO 操作在 LOGO 计算框架下的运行原理。

3.3.1 getSubDataset 操作

getSubDataset 操作用于从 RSP 块组成的 HDFS 文件中采样，这些文件最初是以 HDFS 文件的形式存储在 HDFS 文件系统中，随后通过 RSP 转换算法转换为 RSP 数据块并作为 HDFS 文件存储在 HDFS 文件系统中。通过指定参数 Num，随机从 HDFS 文件中读取 Num 个 RSP 块作为 RspRDD 的分区，用于后续的 LOGO 建模。

3.3.2 LO 操作

LO 操作用来向集群的虚拟机分发一个算法以及相应参数值，算法的输入数据通过 RspRDD 的分区传递。

当一个分布式大数据读入 RDD 后，首先通过数据预处理转换成 RspRDD 的标准格式，这一过程用 Spark 提供的算子完

表 1　一个简单的 RSP-LOGOML 应用程序例子

Table 1　A simple example of RSP-LOGOML application

1	val inputDataset = spark.rspRead.parquet("datas/....")
2	val transactions = inputDataset.getSubDataset(m)
3	val localTable = transactionsRDD. LO(trainDF => logoFPGrowth(support,trainDF))
4	val goTable = localTable.GO(mergeFPGrowth)

成。LO 操作执行算子的输入数据通过 RspRDD 的分区自动传入。Apps 开发者只需设定算子的参数值即可，可以直接选用默许的参数值。

LO 操作是串行算法进行并行计算的桥梁。计算过程中，master 节点将 LO 操作中的任务分发到各个 slave 节点的虚拟机，算法在多个虚拟机中独立并行地执行。在这个过程中，算法在各虚拟机内迭代处理各自的 RSP 样本数据块，得到各自的建模结果。每个结果存在一个输出 RspRDD 的分区，待 GO 操作处理。

3.3.3 GO 操作

GO 操作用于执行一个集成算子，对 LO 操作输出的 RspRDD 的分区结果做集成计算，输出集成结果。对不同类型的算法，RSP-LOGOML 算法库内置了一些集成算子。例如，对有监督学习的分类算子，提供了多数投票集成算子，基于特定的评估指标优选基模型集成算子[16]；对有监督学习的回归算法，提供回归模型的均值集成算子；对无监督学习的 K-Means 聚类算法，提供了基于簇中心平均的集成算子[17]；对分布式频繁项集计算，提供了多数投票集成算子[8]。

除算法库内置集成算子外，可以在应用程序中编写新的集成算子函数，然后在 GO 操作中执行集成算子函数，生成集成结果。用户可以通过集成算子函数灵活地实现不同的集成方法。

3.3.4 Apps 程序例子

一个简单的应用程序例子如表 1 所示。第一行将指定路径的数据集读取为 RspRDD；第二行从数据集中随机采样 m 个 RSP 块，用于后续的流程；第三行将采样的数据按无监督流程转换，得到的 transcationRDD 中的数据格式是输入数组的形式，并使用算法库中的 logoFPGrowth 算法得到多个局部频繁项集结果；第四行使用频繁项集的集成方法得到最后的全局频繁项集结果。

4 实验验证

实验采用了 7 个数据集对 RSP-LOGOML 算法库中的分类、聚类、回归和频繁项集 4 类典型机器学习算法进行了测试，同时也测试了 Spark MLlib 库中的同类算法和单机运行的 Smile 相同算法。算法性能和计算效率实验结果表明，RSP-LOGOML 算法库算法的性能与 Spark MLlib 库和 Smile 相同算法性能相似，但计算效率和数据处理能力显著地提高。

4.1 实验设置

数据集：表 2 给出 7 个数据集的特征。HIGGS 和 MNIST-PCA 数据集是包含 2 类和 10 类对象的真实数据， DS1 是 2 类合成大数据集，用于展示算法处理大数据的能力，这三个数据集用于测试分类和聚类算法。DS2 和 DS3 是人工生成数据，用于测试回归算法。TH2 和 Kaggle 是真实的交易数据集，用于测试频繁项集挖掘算法。

表 2 数据集

Table 2 Datasets

数据集	RSP 块大小	特征维度	记录数	测试任务
HIGGS	64MB	28	11000000	分类、聚类
MNIST-PCA	64MB	86	8100000	分类、聚类
DS1	100MB	20	327360000	分类、聚类
DS2	64MB	20	11000000	回归
DS3	64MB	50	11000000	回归
TH2	19MB	-	26496645	频繁项集
Kaggle	4MB	-	3214874	频繁项集

注：-处指该数据集不存在特征维度。

实验设置：选择了 RSP-LOGOML 中的 11 个分类算法和 3 个聚类算法对 3 个分类和聚类的数据集进行了测试，同时选择了 Spark MLlib 中的 4 个分类算法和 1 个聚类算法，Smile 中的 9 个分类算法和 3 聚类算法做了对比实验。同理，在 RSP-LOGOML 算法库中选了 6 个回归算法对两个回归数据集做了测试，在 Spark MLlib 算法库中选了 3 个同类回归算法和 Smile 中 4 个回归算法做了对比实验。对于两个交易数据集，在 3 个库中选择了 FP-Growth 和 AssocaitonRule mining 算法做了测试。

实验中设定的 RSP 样本数据块的大小如表 2 所示，RSP-LOGOML 算法采用数据

集 10%的数据做近似计算，而 MLlib 和 Smile 算法在整个数据集上做计算。由于计算资源限制，Smile 的有些算法不能完成对一些数据集的计算，而相同算法在 LOGO 框架下不受资源限制。

对于同一算法采用相同的参数设置，具体参数值如表 3 所示。在频繁项集算法实验中，TH2 数据集上的最小支持度设置为 0.02，Kaggle 数据集上的最小支持度设置为 0.002，以确保能够返回足够的频繁项集。

实验环境：分布式计算在 32 个节点的集群上完成，其中 24 节点的配置为两个 Intel Xeon E5-2650 CPU 和 128G 内存，每个 CPU 有 16 个核，频率为 2.6GHz。其余 8 个节点配置为两个 Intel Xeon E5-2630 CPU 和 128G 内存，每个 CPU 有 12 个核，频率为 2.6GHz。软件为 Spark 2.4.0 和 YARN 3.0.0。运行 Spark MLlib 和 RSP-LOGOML 算法时，配置了 20 个 executor 进程，每个 executor 进程分配 16G 内存资源。Smile 算法单机运行在 master 节点的服务器上完成，每个算法配置 48G 内存。

表 3　算法参数设置

Table 3　Parameter setting of different algorithms

算法	参数设置值
逻辑回归	maxIter=500, regParam=0.1
决策树	maxDepth=10, nodeSize=10
随机森林	ntrees=10, maxDepth=10, nodeSize=10
支持向量机	regParam=0.1, tol=1.00E-06
线性回归	-
回归树	maxDepth=10, nodeSize=10
回归森林	ntrees=10, maxDepth=10, nodeSize=10
K-均值	maxIter=100, tol=1.00E-04
FPGrowth	support(Kaggle)=0.002, support(TH2)=0.02

注：- 处指算法不需设置参数。

4.2　评估指标

运行时间：我们用 RSP-LOGOML、Spark MLlib 和 Smile 算法的运行时间度量计算效率。RSP-LOGOML 算法的运行时间记为从 LO 操作开始到 GO 操作结束的时间；对于 Spark MLlib 和 Smile 算法的运行时间记为算法开始运行到结束的时间。

性能：分类算法用分类精度评估性能，计算公式为：

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

其中，TP 为正类正确分类的样本数，FP 为正类错误分类的样本数，TN 为负类正确分类的样本数，FN 为负类错误分类的样本数。

回归算法性能的计算公式为：

$$R2 = 1 - \frac{SSE}{SST} \quad (2)$$

其中，SSR 表示回归平方和，SST 表示总平方和。

聚类算法用纯度作为性能度量，计算公式为：

$$Purity = \frac{1}{N}\sum_{1}^{k} \max_{i \in \{1,\cdots,I\}} |\omega_k \cap c_i| \quad (3)$$

其中，N 为样本总数，ω_k 是第 k 个聚类集合，c_i 是第 i 个真实类别，$|\omega_k \cap c_i|$ 表示聚类集合 ω_k 和真实类别 c_i 的交集的样本数量。

由于 RSP-LOGOML 算法库中频繁项集算法采用近似计算方法，以 Spark MLlib 或 Smile 的算法计算结果为真值，算法性能按 Sun 等提出的分布式集成方法[8]，性能评估采用精确度计算公式

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \quad (4)$$

和召回率的计算公式

$$\mathrm{Re}call = \frac{TP}{TP + FN}。 \quad (5)$$

其中，TP 为 RSP-LOGOML 算法计算结果中为真的频繁项集的数量，FP 为真的频繁项集没有出现在 RSP-LOGOML 计算结果的数量，FN 为频繁项在 RSP-LOGOML 的计算结果中预测为非频繁项的数量。

4.3　实验结果

4.3.1 分类、聚类实验结果

在 RSP-LOGOML 中选用了 11 个分类算法和 3 个聚类算法对 HIGGS 数据集进行了测试，结果见表 4。算法名在最左列显示，最后 3 个算法为聚类算法。RSP-LOGOML、Spark MLlib 和 Smile 下面的两列为算法运行时间和性能。“-”表示 Spark MLlib 中不存在相应算法，“×”表示相应算法由于资源限制不能产生最后结果。

先看前 4 个分类算法的运行时间，RSP-LOGOML 算法的运行时间比 Spark MLlib 和 Smile 的算法要快两倍以上。其中，Smile 的支持向量机算法没有产生最后结果。再看算法性能，RSP-LOGOML 中逻辑回归和支持向量机的模型精度高于 Spark MLlib 同类模型的精度，因为 RSP-LOGOML 算法产生的是集成模型。三个库中其它两个算法的性能结果相同，但 RSP-LOGOML 算法的模型只用了 10%的数据，其它两个库的算法用全部数据。

另外 7 个分类算法只能比较 RSP-LOGOML 和 Smile 算法的结果，可以看出，两个库中的算法模型精度基本相同，除了判别分析算法的计算时间近似外，其它复杂算法，RSP-LOGOML 要比 Smile 算法的计算时间短的多。

对于聚类算法，由于 Spark MLlib 中只有 K-均值算法做对比。三个库中的算法聚类性能基本相同，但 RSP-LOGOML 算法的运行时间要比其它两个库中算法的运行时间快的多。

表 5 给出 MNIST-PCA 数据集的测试结果。由于这个数据集有 10 个类别，不能采用二分类的算法，表中只列出 7 个分类算法的结果。前三个算法的结果可以看出，RSP-LOGOML 算法的集成模型好于 Spark MLlib 和 Smile 算法的单个模型。Spark MLlib 的决策树算法略好于 RSP-LOGOML 的决策树算法，因为后者只用了 10%的数据。从运算时间上看，RSP-LOGOML 算法要远快于其它两个库的算法。

对于其它分类算法，Smile 的径向基函数网络无法处理 MNIST-PCA 数据集，另外 3 个分类算法，RSP-LOGOML 算法的模型精度和计算时间都优于 Smile 的同类算法。

对于聚类算法，Smile 的 K-均值、G-均值、X-均值算法由于内存资源不足均无法处理此数据集，Spark MLlib 只有 K-均值可用，聚类性能与 RSP-LOGOML 的 K-均值相似，但运行时间长的多。

表 6 为人工合成的 DS1 数据集的实验结果，由于 DS1 数据很大，Smile 单机建模无法进行，所以表中没有 Smile 的实验结果。

从运行时间上看，RSP-LOGOML 算法比 Spark MLlib 算法运算快的多。对比较小的 HIGGS 和 MNIST-PCA 数据集，RSP-LOGOML 算法在这个大数据集上计算效率优势更为显著。在性能上，RSP-LOGOML 算法结果与 Spark MLlib 算法结果相当接近。4.3.2 回归实验结果

表 7 和表 8 给出 6 个回归算法在数据集 DS2 和 DS3 上的实验结果。Spark MLlib 只有 3 个回归算法可用。从性能上看，这三个库的回归算法性能相似，但 RSP-LOGOML 算法只用 10%的数据建模，而其它两个库的算法用全部数据建模。在运算时间上，RSP-LOGOML 的时间性能更优。由于运算的复杂度，相比其它算法，Smile 的支持向量机和 Lasso 回归在单机上处理这两个数据集需要大量的时间。

4.3.3 频繁项集挖掘实验结果

表 4　HIGGS 数据集的实验结果

Table 4　Experimental results on HIGGS dataset

算法	RSP-LOGOML		Spark MLlib		Smile	
	时间(s)	性能	时间(s)	性能	时间(s)	性能
逻辑回归	25.6957	0.6438	94.1230	0.6227	51.4742	0.6409
支持向量机	455.4721	0.6428	2125.0173	0.5934	×	×
决策树	15.2526	0.7025	111.2849	0.7048	120.3926	0.7007
随机森林	73.5483	0.7061	141.2220	0.7048	238.4653	0.7025
线性判别分析	16.9560	0.6427	-	-	18.3749	0.6444
正则判别分析	18.1460	0.6497	-	-	33.8599	0.6472
二次判别分析	36.5683,	0.6451	-	-	17.0502	0.6444
Fisher 线性判别	17.8329	0.6397	-	-	19.4387	0.6367
多层感知器神经网络	105.1328	0.5254	-	-	2322.0648	0.5279
径向基函数网络	25.8267	0.5277	-	-	×	×
自适应提升算法	228.7279	0.7170	-	-	13759.99	0.7106

K-均值	39.9498	0.5003	101.5477	0.5014	165.6183	0.5013
G-均值	165.0775	0.5001	-	-	333.9994	0.5013
X-均值	67.1201	0.5001	-	-	374.4661	0.5013

注：×处指算法缺乏足够的资源运行，-处指 Spark MLlib 不存在该算法。

表 5　MNIST-PCA 数据集的实验结果

Table 5　Experimental results on MNIST-PCA dataset

算法	RSP-LOGOML		Spark MLlib		Smile	
	时间(s)	性能	时间(s)	性能	时间(s)	性能
逻辑回归	88.8204	0.8621	119.1110	0.8018	823.6211	0.8620
决策树	51.5007	0.7130	111.2849	0.7216	278.8387	0.6932
随机森林	107.3951	0.8330	366.5287	0.8168	637.4727	0.7870
线性判别分析	58.4430	0.7781	-	-	100.8224	0.7781
Fisher 线性判别	33.1034	0.8005	-	-	99.9653	0.7995
径向基函数网络	37.7638	0.4991	-	-	×	×
自适应提升算法	342.6394	0.8394	-	-	20793.259	0.8065
K-均值	68.4321	0.5292	134.4077	0.5330	×	×
G-均值	135.2492	0.5232	-	-	×	×
X-均值	102.0382	0.4951	-	-	×	×

注：×处指算法缺乏足够的资源运行，-处指 Spark MLlib 不存在该算法。

表 6　DS1 数据集的实验结果

Table 6　Experimental results on DS1 dataset

算法	RSP-LOGOML		Spark MLlib	
	时间(s)	性能	时间(s)	性能
逻辑回归	76.3670	0.9135	1038.0678	0.9135
支持向量机	2418.0253	0.9138	10323.5590	0.9136
决策树	133.3115	0.8832	553.6867	0.8761
随机森林	212.4092	0.8856	1076.2559	0.8893
K-均值	64.6130	0.9135	1040.5700	0.9137

表 7　DS2 数据集的实验结果

Table 7　Experimental results on DS2 dataset

算法	RSP-LOGOML		Spark MLlib		Smile	
	时间(s)	性能	时间(s)	性能	时间(s)	性能
线性回归	16.2989	0.9999	94.1158	0.9999	41.9044	0.9999
回归树	41.1402	0.8205	109.3965	0.7984	107.5136	0.7655
回归森林	46.4018	0.8443	141.2975	0.8302	202.0270	0.8075
支持向量机（回归）	385.1803	0.9990	-	-	49264.3644	0.9993
岭回归	49.6568	0.9999	-	-	24.4567	0.9999
Lasso 回归	60.8375	0.9999	-	-	244.5691	0.9999

注：-处指 Spark MLlib 不存在该算法。

表 8　DS3 数据集的实验结果

Table　8 Experimental results on DS3 dataset

算法	RSP-LOGOML		Spark MLlib		Smile	
	时间(s)	性能	时间(s)	性能	时间(s)	性能
线性回归	28.4118	0.9997	146.8596	0.9997	122.3382	0.9999
回归树	93.1663	0.7597	160.8822	0.7084	254.9991	0.7655
回归森林	80.5679	0.7829	183.9823	0.7560	440.2951	0.7261
支持向量机（回归）	2167.53	0.9993	-	-	131108.8015	0.9995
岭回归	68.5597	0.9997	-	-	98.5261	0.9997
Lasso 回归	71.2883	0.9997	-	-	1402.9357	0.9998

注：-处指 Spark MLlib 不存在该算法。

用　FP-Growth　和　AssocaitonRule　Mining

算法对交易数据集 TH2 和 Kaggle 进行了测
试，TH2 数据集的测试结果如表 9 和表 10 所示，FPG 表示 FP-Growth, ARM 表示 AssocaitonRule Mining，性能用 Precision 和 Recall 表示，运行时间用秒表示。

首先运行 Spark MLlib 的 FP-Growth 和 Smile 的 FP-Growth 与 ARM 算法计算数据集的频繁项集作为真值，再运行 RSP-LOGOML 的 FP-Growth 和 ARM 计算近似的频繁项集，用公式（4）和（5）计算 Precision 和 Recall。两个算法的 Precision 和 Recall 结果如表 9 所示，可以看出，两个算法都得到 100%的召回率，说明可以得到所有真实的频繁项集。精度也都大于 97%，满足实际应用要求。

表 10 所示运算实际显示，RSP-LOGOML 算法的计算效率比 Spark MLlib 算法稍有优势，比 Smile 单机算法具有明显的优势。

表 11 和表 12 展示了 Kaggle 数据集的实验结果。RSP-LOGOML 算法同样有很高的召回率和 90%的精度。与 TH2 数据集结果的差异是 Kaggle 数据集较小，设置的 RSP 样本数据快也较小，造成算法精度下降，同时，分布式计算的时间也长于单机计算时间。这个结果指出，分布式计算的优势在于处理大数据，特别是单机无法处理的数据。在小数据集计算上不需要分布式计算。

4.4 讨论与分析

Smile 串行算法库的优点是包含丰富的算法供数据分析时选用，但数据处理能力受单机计算资源的限制。随着数据集的不断增大，单机串行算法分析模式已经难以满足大数据分析的需求。

分布式并行计算是大数据分析的必要手段。但是，基于谷歌 MapReduce 编程模型和计算框架的大数据分布式并行计算存在在如下瓶颈：（1）由于复杂串行算法难以用 MapReduce 编程模型改写，算法库中可用的分布式并行计算算法不足；（2）MapReduce 计算框架在执行迭代算法时，由于节点间的数据通信开销，算法执行效率低下，数据处理能力受内存约束，大数据扩展性差。

LOGO 计算框架从根本上消除了 MapReduce 计算瓶颈，开辟了串行算法无需改写、直接高效执行分布式并行计算的新模式，极大地丰富了分布式并行计算的算法库。

RSP-LOGOML 算法库是大数据分布式计算的一种新技术。本文展示的实验结果反映，对于较大的数据集，RSP-LOGOML 算法库中算法的计算效率远高于同类算法的单机计算效率和 Spark MLlib 中 MapReduce 编程同类算法的计算效率。RSP-LOGOML 算法库中算法的性能等价或略好于 Spark MLlib 中的算法，但可用算法的数量远大于 Spark MLlib。因此，RSP- LOGOML 算法库为数据分析科学家提供了更好和更多的算法工具。

分布式并行计算的优势在于处理大规模数据集。对于小数据，单机串行算法的

表 9　TH2 数据集的性能结果

Table 9　Evaluation results on TH2 dataset

算法	Precision	Recall
FPG	0.9772	1
ARM	0.9845	1

表 10　TH2 数据集各算法的运行时间

Table 10　Runing time of different algorithms on TH2 dataset

算法	RSP-LOGOML	Spark MLlib	Smile
	时间(s)	时间(s)	时间(s)
FPG	53.831195	72.515438	459.6327
ARM	48.263593	-	528.1199

注：-处指 Spark MLlib 不存在该算法

表 11　Kaggle 数据集的性能结果

Table 11　Evaluation results on Kaggle dataset

算法	Precision	Recall
FPG	0.9003	0.9966
ARM	0.8962	0.9949

表 12　Kaggle 数据集各算法的运行时间

Table 12　Runing time of different algorithms on Kaggle dataset

算法	RSP-LOGOML	Spark MLlib	Smile
	时间(s)	时间(s)	时间(s)
FPG	25.9255	66.2956	18.4779
ARM	26.2764	-	16.2963

注：-处指 Spark MLlib 不存在该算法

计算效率更高，因为分布式并行计算会产生节点间的通信开销。这一点可以从表 12 中 Kaggle 数据集的结果中看出。因此，单机串行算法库在小数据分析中优势明显，而 RSP-LOGOML 算法库将单机串行算法库移植到分布式计算平台上，为大数据的处理与分析提供了更加丰富的算法资源。

5 结 论

本文介绍了 LOGO 分布式计算框架下的机器学习算法库 RSP-LOGOML。LOGO 采用的是 Non-MapReduce 的计算范式，消除了迭代算法运行时节点之间的数据通信，在 RSP 分布式数据表达模型支持下，实现了串行算法在 LOGO 框架下的直接分布式运行，丰富了大数据分布式并行计算的算法工具。

RSP-LOGOML 算法库是对大数据分布式并行计算的框架模型创新和分析工具创新。依据的近似计算方法、LOGO 计算框架和 RSP 数据表达模型，都是近年出现的原创性成果。笔者下一步工作将采用更大的数据集以及真实应用的数据集对 RSP-LOGOML 算法库的性能进行测试，为大数据分析提供更丰富、效率更高的算法库。

参考文献

[1] Salloum S, Dautov R, Chen X, Peng P X, Huang J Z. Big data analytics on Apache Spark[J]. Int. Journal. Data Science and Analytics, 2016, 1(3-4): 145-164.

[2] Sun X, He Y, Wu D, et al. Survey of Distributed Computing Frameworks for Supporting Big Data Analysis[J]. Big Data Mining and Analytics, 2023, 6(2):154-169.

[3] Sun X, Zhao L, Chen J, et al. Non-MapReduce computing for intelligent big data analysis[J]. Engineering Applications of Artificial Intelligence, 2024, 129: 107648.

[4] Salloum S, Huang J Z, He Y. Random sample partition: a distributed data model for big data analysis[J]. IEEE Transactions on Industrial Informatics, 2019, 15(11): 5846-5854.

[5] 黄哲学, 何玉林, 魏丞昊,等. 大数据随机样本划分模型及相关分析计算技术[J]. 数据采集与处理, 2019, 034(003):373-385.
Huang ZX, He YL, Wei CH et al. Random Sample Partition Data Model and Related Technologies for Big Data Analysis [J]. Journal of Data Acquisition and Processing, 2019, 034(003):373-385.

[6] Mahmud M S, Huang J Z, Salloum S, et al. A survey of data partitioning and sampling methods to support big data analysis[J]. Big Data Mining and Analytics, 2020, 3(2): 85-101.

[7] Wei C, Salloum S, Emara T Z, et al. A two-stage data processing algorithm to generate random sample partitions for big data analysis[C]//Cloud Computing – CLOUD 2018. LNCS 10967, Springer, 2018: 347-364.

[8] Sun X, Ngueilbaye A, Luo K, et al. A scalable and flexible basket analysis system for big transaction data in Spark[J]. Information Processing & Management, 2024, 61(2): 103577.

[9] Thakur B S, Bansal K L. Performance Evaluation Of Apache Hadoop, Apache Spark, And Apache Flink[J]. Advances In Management, Social Sciences and Technology, 2021: 93-100.

[10] Sharma A, Puri D, Kumar M, et al. Implementation and comparison of big data analysis on large dataset using spark and flink[C]//ICCCE 2021: Proceedings of the 4th International Conference on Communications and Cyber Physical Engineering. Singapore: Springer Nature Singapore, 2022: 385-394.

[11] Hueske F, Kalavri V. Stream processing with Apache Flink: fundamentals, implementation, and operation of streaming a

pplications[M]. O'Reilly Media, 2019.
（评审意见 1 和 2 的第 2 点）
[12] Luna J M, Padillo F, Pechenizkiy M, et al. Apriori Versions Based on MapReduce for Mining Frequent Patterns on Big Data[J]. IEEE Transactions on Cybernetics, 2018, 48(10): 2851-2865.
[13] Assefi M, Behravesh E, Liu G, et al. Big Data Machine Learning using Apache Spark MLlib[C]//IEEE Big Data 2017. DOI:10.1109/BigData.2017.8258338.
[14] Meng X, Bradley J, Yavuz B, et al. Mllib: Machine Learning in Apache Spark [J].The journal of machine learning research, 2016, 17(1): 1235-1241.
[15] Boden C, Spina A, Rabl T, et al. Benchmarking data flow systems for scalable machine learning[C]//Proceedings of the 4th ACM SIGMOD Workshop on Algorithms and Systems for MapReduce and Beyond. 2017: 1-10.
[16] Siegal D, Jia G, Agrawal G. Smart-MLlib: A High-Performance Machine-Learning Library[C]// IEEE International Conference on Cluster Computing. 2016: **D**OI: 10.1109/CLUSTER.2016.49
[17] Cid-Fuentes J A, Sola S, Alvarez P, et al. dislib: Large Scale High Performance Machine Learning in Python[C]//The 15th International Conference on eScience. Barcelona, Spain, 2019.
[18] 徐继伟,杨云.集成学习方法:研究综述[J].云南大学学报(自然科学版),2018,40(06):1082-1092.
Xu J W, Yang Y. A survey of ensemble learning approaches [J]. Journal of Yunnan University (Nature Science Edition), 2018,40(06):1082-1092.
[19] 张春霞, 张讲社. 选择性集成学习算法综述[J].计算机学报, 2011, 34(08):1399-1410.
Zhang C X, Zhang J S. A survey of selective ensemble learning algorithms [J]. Chinese Journal of Computers, 2011, 34(08):1399-1410.
[20] 吉根林,凌霄汉,杨明.一种基于集成学习的分布式聚类算法[J].东南大学学报(自然科学版),2007,(04):585-588.
Ji G L, Lin X H, Yang M. Distributed clustering algorithm based on ensemble learning [J]. Journal of Southeast University (Nature Science Edition),2007,(04):585-588.
[21] Diwakar T, Kumar S A, Ramachandra R B, et al. Credit Scoring Models Using Ensemble Learning and Classification Approaches: A Comprehensive Survey[J]. Wireless personal communications: An Internaional Journal, 2022, 123:785-812.
[22] Mienye I D, Sun Y. A survey of ensemble learning: Concepts, algorithms, applications, and prospects[J]. IEEE Access, 2022, 10: 99129-99149.
[23] Dong X, Yu Z, Cao W, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14: 241-258.
[24] Campagner A, Ciucci D, Cabitza F. Aggregation models in ensemble learning: A large-scale comparison[J]. Information Fusion, 2023, 90: 241-252.
[25] Mahmud M S, Huang J Z, Garc á S. Clustering approximation via a fusion of multiple random samples[J]. Information Fusion, 2024, 101(C): doi.org/10.1016/j.inffus.2023.101986
[26] Mahmud M S, Huang J Z, Ruby R, et al. Approximate Clustering Ensemble Method for Big Data[J]. IEEE Transactions on Big Data, 2023, 9(4):1142-1155.
[27] Minaei-Bidgoli B, Topchy A P, Punch W F. A comparison of resampling methods for clustering ensembles[C]//IC-AI. 2004: 939-945.
[28] Mahmud M S, Huang J Z, Ruby R, et al. An ensemble method for estimating the number of clusters in a big data set using multiple random samples[J]. Journal of Big Data, 2023, 10(1): 1-33.

[29] Xiao W, Hu J. Paradigm and Performance Analysis of Distributed Frequent Itemset Mining Algorithms Based on MapReduce[J]. Microprocessors and Microsystems, 2021:doi.org/10.1016/j.micpro.2020.103817.