文章编号：2095-2163(2020)06-0148-05 中图分类号：TP311 文献标志码：A

面向 MapReduce 的大数据分类模型及算法

柯建波
（广东工业大学 华立学院，广州 511325）

摘 要：针对传统的大数据分类模型及算法存在处理数据时间长的缺陷问题，开展了面向 MapReduce 的大数据分类模型及算法的研究。构建大数据分类模型，挖掘有效数据点，排列数据组合方式，划分数据局部连接方式，获取局部节点微簇数据，计算节点数据组中冗余数据及无效数据增量值，重构中心节点样本算法，调整集成数据分类策略，优化更新数据集成分类方式。设计实验，模拟实验环境，验证提出模型计算法在实际应用中可缩短数据处理时间，具有实际研究价值。

关键词：MapReduce；大数据；分类模型；算法研究

Big data classification model and algorithm for MapReduce

KE Jianbo
（Huali College, Guangdong University of Technology, Guangzhou 511325, China）

【Abstract】 The traditional big data classification model and algorithm have the defect of long processing time, so the research on big data classification model and algorithm for MapReduce is carried out. Construct big data classification model, mine effective data points, arrange data combination mode, divide data local connection mode, obtain local node micro-cluster data, calculate redundant data and invalid data increment in node data set, reconstruct central node sample algorithm, adjust integrated data classification strategy, optimize and update data integration classification mode. Design the experiment, simulate the experimental environment, and verify that the proposed model calculation method can shorten the data processing time in practical application, which has practical research value.

【Key words】 MapReduce; big data; classification model; algorithm research

0 引 言

MapReduce 作为一种大型互联网编译模型，主要用于实施大规模的数据聚类并行计算（数据存储空间超过 1TB），数据概念映射方式及黑盒解题思路是编译程序的主要提出方式，基于数据函数编译方式及多种计算机编译语言，在使用中可依照矢量编译语言特性，提供计算机编程人员分布并行计算模式，有关计算指令可在计算机语言的调试作用下并行输出[1]。目前使用该技术实现主要是指结合数据间的联系性，设计合理的映射类型函数，建立数据集合链接，将单个键位组织通过映射联系组合成新的计算机键位对，确保键位对中计算数值的每一个共享键位值相同。结合大数据技术目前在市场的应用情况，大数据分类模型理念最早提出于 20 世纪 80 年代初，市场强大的应用需求使分类模型的研究成为技术调研重点。

大数据技术在逐步发展中已经上升至国家层面，因此，本文将提出面向 MapReduce 的大数据分类模型及算法的研究。

1 面向 MapReduce 的大数据分类模型及算法

将调研重点聚焦在具有流动特征数据层面上，以传统数据挖掘技术为基础，采样收集单一样本的方式，提出数据合理的学习方法，根据数据流及特征数据集的不同显示方法，存储离线式挖掘数据，由于无法一次性完全处理数据，因此在数据处理前期应对数据实施聚类管理，区分静态数据与动态数据，依照数据的表现形式，创新数据集成理念，引入新型数据分类技术，探索分布数、数据传递方法，使用检索、汇聚、连接、分离、清洗等方式优化对应算法，提供数据分类更加优化模式。

1.1 构建面向 MapReduce 的大数据分类模型

给定处理数据流 T 及数据分类标识集合 C，合理选择数据分类器，明确数据分类法则，描述数据分类过程，动态收集依照时间变化的数据发展趋势，强调数据处理中选择数据的质量，完整正向数据集合与负向数据集合分类，提出数据抽样检测技术，构建面向 MapReduce 的大数据分类模型[2]，如图 1 所示。

根据上述图 1 所述信息，设定数据流经整体时间为 t，输入数据点为数据采集点，经过一个时间节点，在此收集数据，为挖掘数据有效点，表示为 $t-1$。

作者简介：柯建波（1981-），男，学士，讲师，主要研究方向：人工智能、计算机网络。

收稿日期：2020-02-25

由数据中心服务器提供数据当前分类设备及此前数据集成设备，遵循数据流经设备的运行模式，输出节点数据，表示为 t，引入数据挖掘技术处理当前状态下时间点。定义数据处理模式为 M，划分数据处理模块。提取历史窗口数据中局部数据，定义上次处理数据模式为 $M(T-1)$，当前数据局部处理模式为 $M(T)$，输出一次处理数据，按照数据特征划分数据集合[3]。引入 MapReduce 分布式局部数据处理方式，定义中间界数据为集合中样本训练集，并将该部分数据归为网络监控数据，整合一次数据集，获取二次数据处理方式。根据获取数据时间点的增多，训练使用的样本数据集合处理模式同步发生改变，数据处理模式用 Chunk 表示，取值为 1 ~ n 之间任意实数。依照数据集合的不断汇聚，调整潜在的数据学习模式，设定数据挖掘目标，调整数据时间变化模式，以单元为模块划分数据集合[4]。定义时间序列参数为 T，则单元数据时间序列表达方式为 $T=\{t_1,\cdots,t_n\}$，定义数据流表达形式为 S，则数据流中具体数据集合表达形式为 $S=\{r_1,\cdots,r_n\}$。K 表示为流经数据的中心节点，整理数据模式，处理历史窗口数据集合，则数据单元中任一数据即可近似看作数据处理方式，提取数据中特征点数据，统一数据格式，结合大数据处理方式，归入指定数据库，完成面向 MapReduce 的大数据分类模式构建。

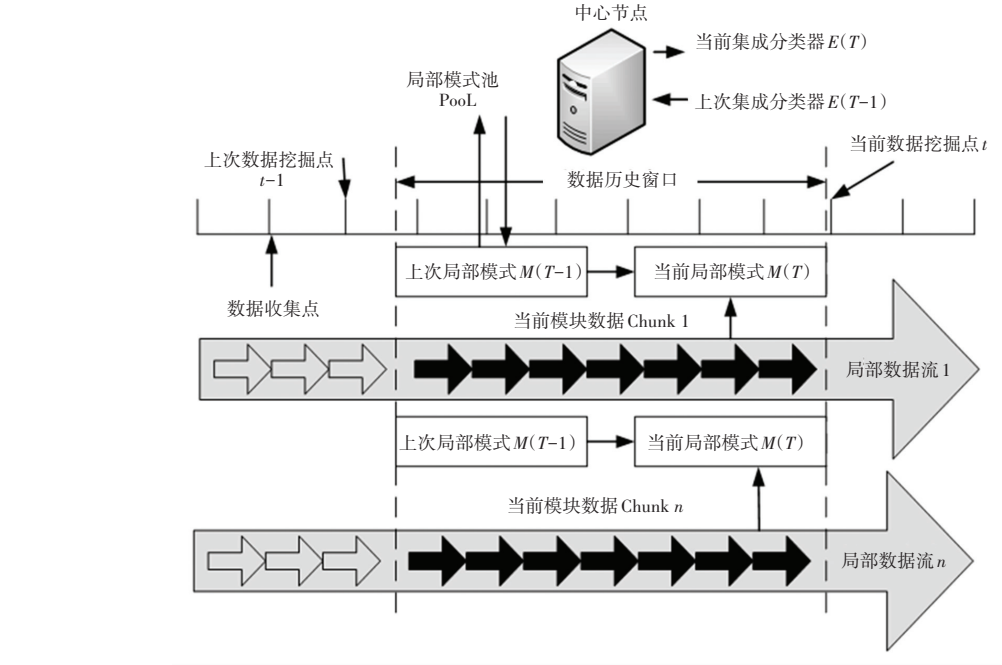

图 1　面向 MapReduce 的大数据分类模型

Fig. 1　MapReduce-oriented big data classification model

1.2　获取局部节点微簇数据

依照上述提出的大数据分类模型，整合有理数据，绘制成具有自身独立特征的数据聚类集合，合理选择数据聚集方式，计算分布式数据组的数据总和、集合中极值数据、代数数据与整体数据表达模式[5]。设定局部节点中分支数据表达方式为 n，局部数据算法表达如下公式所示：

$$M,c=\left(\sum_{i=1}^{n}x_i^j\right)/n. \tag{1}$$

式中，M 表示为获取局部数据算法；c 表示为选定样本数据集合；x 表示为局部数据传递方式；n 表示为局部节点中分支数据表达方式；i 表示为挖掘数据，公式中用单元“1”表达；j 表示为挖掘数据终端输出模式。根据上述计算公式，可将分类数据集合按照数据划分依据分成分布式数据集合、代数数据集合、整数数据集合 3 种[6]。在每个计算单元中均可获取具有特征值的数据节点，计算数据节点的微簇数据。如下公式所示。

$$N,b=\sqrt{\int_y y_j^n}\,(d*w\lambda). \tag{2}$$

式中，N 表示为局部节点的微簇数据表达方式；b 表示为局部微簇样本数据集合；y 表示为样本中有效数据集；d 表示为数据流经历史窗口实际时间；w 表示微簇数据获取误差值，0.25 ~ 0.65 为数据组可调节范围；λ 表示为数据获取时间参数值，通常情况下取值在 0.3 ~ 0.45 之间。使用上述计算公式，可直接获取局部数据集合中簇组数据，按照多元代数函数计算方式，每个局部数据组的参数表达方法，即可

按照数据标准表达方式输出终端获取数据值[7]。上述计算中涉及的数据值均为数据分类模型中数据组,可按照基本算法流程,采用近似值表达方式将簇组数据以常规方式表达。

1.3 计算节点数据增量

根据上述获取的局部节点微簇数据,连接两个或两个以上数据组,按照标准记录条件将其划分成 TPC-H 数据集,结合数据查询用例,运用多种计算方式,叠加数据组,筛选数据重合部分,将其统一格式后纳入数据库中管理。清洗数据中特征点,去除重叠数据,将待处理数据组中冗余或无关数组值去除,提升剩余数据值质量,制定格式实施数据转换,逐条处理待处理信息[8]。引入 Map 任务处理模式,计算数据组中冗余数据增值。计算公式如下:

$$\mathrm{dis}(p_2 - p_1) = \sqrt{\sum_{i}^{H} (p_1 - p)^2}. \tag{3}$$

式中, dis 表示为加强数据项; p 表示为样本数据总值; p_1 表示为重叠数据组; H 表示为数据清洗模式; p_2 表示为带解析数据组。根据上述计算公式,提出数据处理次数,结合数据处理执行时间 T 与数字/模拟转换次数的函数关系,分析处理数据检索方式。对照 MapReduce 技术,增加节点数据个数,提高数据执行效率。

在数据排列阶段,按照组间合成数据值实施数据聚类操作,组合具有相同键位的数据值,计算对数据综合值平方差,得到终端输出数据值单个组值的频次值。因为数据组中包含个别极端数据[9]。因此,应按照原始数据的恢复性能,统计微簇数据组的原始数据值,由于数据值中统计的组合数据比输出数据的抽象值更高,可采用引入 C5.4 计算模式,基于全局统计方法,提出数据组的灰度计算方法,分析特征数据点的灰度值,更加有利于推进整体算法。

1.4 中心节点样本算法重构

结合上述计算的数据增量,将中心节点样本重构分成主要 3 个独立步骤:(1)局部挖掘数据重构,定义每个数据组中节点,按照数据中心点划分数据应用模块,收集中心点附近数据,整理成数据集合,按照单元定义模式维护早期设定的数据挖掘点,形成新的数据增量集合,构建全新算法微簇数据处理方式[10-11]。(2)根据多个局部数据组,按照数据不同传递模式,更新整理数据中心节点,连接互联网,更新完成后通过数据互联网传递方式,将多个中心节点数据传送至整体数据集中心节点中。(3)引入全局数据挖掘模式,整合数据学习方式,更新数据所属状态。对其中任意一个样本数据实施根距离计算。计算公式如下:

$$u_j^i = \frac{\sum_{i==1}^{n} \{d^{(i)} = j\}\, r^{(i)}}{\{d^{(i)} = j\}}. \tag{4}$$

式中, u 表示为样本两点数据之间的根距离; d 表示为数据自身长度值; j 表示为数据增量值; r 表示为数据重构模式; i 表示为数据组数量。通过上述计算,可重置数据组中心节点,在数据迭代终止过程中,若数据组中心节点位置不发生数据唯一,表明数据重构终止,可输出节点中心位置。反之,将数据组返回上述重构步骤,二次重构数据组,直至数据中心节点与重构数据中心节点重合。

1.5 集成分类更新算法优化

选择数据基础较弱的数据分类装置,采用 C5.4 计算方式,隔离多个局部数据组,基于数据分类装置的优化机制,调整集成数据分类策略。按照数据组权重比值,采用决策树处理数据方式,降低模型及设备对数据处理的干扰性。将满足优化的数据按照数据库标准行的方式连接,比较多种集成分类算法的优缺点。如表 1 所示。

表 1 集成分类算法计算比较

Tab. 1 Comparison of integrated classification algorithms

算法名称	计算模型	权重值	语言表达	用途
Hadoop	读出数据	重	Java	密集数据计算
Mars	读出数据	轻	C++	多次数据预处理
Disco	读出数据	重	C#	大数据处理模式
Haloop	读出数据	重	C#/C++	支持数据迭代计算
Twister	读出数据	轻	Python	增加数据计算终止条件

根据上述表 1 中所述信息,整合多种算法计算方式用途,优化大数据分类模型计算法,采用自然连接的方式,将默认数据值按照权重值排列,组合笛卡尔连接方式,以全连接、半连接等方式,将满足优化的算法实施等值连接,实现集成分类算法的更新及优化。

2 实 验

2.1 实验准备

提出实验,验证本文设计的面向 MapReduce 的大数据分类模型及算法具备一定研究机制,引入 CPU(KDD)公共数据处理技术,搭建大数据挖掘检测数据集合。实验需准备 3 台计算机设备,一台计算机设备为实验主机,设定主机为数据通过主节点(Master),剩余两台计算机设备匹配子节点数据。设备具体属性值如下:处理器选择因特尔(R)核心(TM),i7-5600 运行处理模式;计算机硬盘内存为 64GB;外设硬盘运行内存为 256 GB;计算机运行系

统版本为 Ubuntu15.6；JAVA 计算机语言包工具运行版本为 2.6；集散式系统基础运行框架版本为 2.6.4；仿真实验运行环境为集成式开发运行环境，配备 MapReduce 数据插件，计算机数据集处理语言选择 java。遵循标准测试数据库中数据集，设定该数据集合中共有 60 万个数据样本，样本中包含 55 个不同数据属性变量值，依照数据类别划分为 8 种，占用计算及运行内存 79.5 MB。此次实验从 60 万个数据样本中，随机选取部分数据集合作为此次实验的对照组实验数据，同时选取同样数值的数据集合作为实验组测试数据，为提升实验结果的真实性，两组数据中不可包含重复数据。设定 10 组实验数据，数据选取具体情况如表 2 所示。

表 2　仿真实验数据选取

Tab. 2　Selection of simulation experiment data

次数	对照组数据/例	实验组数据/例	备注
1	13 562	14 590	
2	56 821	57 239	
3	125 890	127 461	
4	163 674	168 946	数据安全，每组选择数据中无重复数据值
5	196 406	198 701	
6	214 535	210 525	
7	217 435	214 524	
8	226 511	222 705	
9	238 520	235 215	
10	235 065	232 726	

依照上述提出的运行实验数据及实验运行环境，忽略其它影响实验结果的外界因素。先采用传统的大数据分类模型及算法，按照上述提出的数据集合，实施数据处理并分类，定义该组为实验的对照组。再采用本文设计的面向 MapReduce 的大数据分类模型及算法实施相同步骤的操作，定义该组实验组。

2.2　实验结果分析

输出实验结果，整理实验中产生的实验数据，绘制成曲线图，如下图 2 所示。

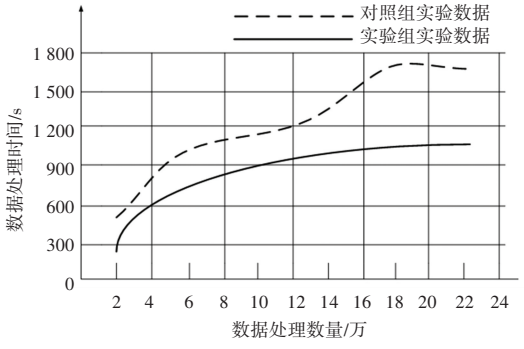

图 2　实验结果

Fig. 2　Experimental results

根据实验过程及实验中产生的实验数据，可得出下述实验结论：随着样本数据量的提升，提出算法处理数据时间平稳上升，具有一定的函数规律，且达到时间峰值后相对平稳。传统方法数据处理时间较不稳定，且处理时间上升速度较快，无明显规律。因此，相比传统的大数据分类模型及算法，本文设计的面向 MapReduce 的大数据分类模型及算法，在实际应用中可有效缩短数据处理时间。弥补了传统算法中针对大量数据集时数据量不足的缺陷，有效地提升了计算效率，提高了大数据分类模型的运行速度，具有实际应用价值。

3　结束语

随着数据分类技术在市场的广泛应用及大数据处理技术的不断更新，本文提出了面向 MapReduce 的大数据分类模型及算法的研究。设计实验，模拟实验环境及实验数据，验证本文设计算法在实际应用中可有效地缩短数据处理时间。尽管本文研究已经趋近于完善，但在实际应用中没有针对数据的迭代情况开展详细分析，因此，在后期的发展中，将基于大数据技术，将数据组按照分类模式及标准误差处理方法，对数据整理实施全方面的优化，提供数据处理终端云平台，模拟数据可能出现分类误差的多种情况，根据可能出现的现象，调整算法的数据分类方式，优化数据外化内存，从多个角度考虑影响数据表达因素，进而为大数据分类模型及算法的研究提供数据支撑。

参考文献

[1] 廖寒逊，滕欢，卢光辉. 基于 MapReduce 的电力大数据增量式属性约简方法[J]. 电力系统自动化，2019(15)：186-192.
[2] 邓小盾. 一种基于大数据的网络日志分析模型构建研究[J]. 电子设计工程，2017，25(23)：97-100.
[3] 翟俊海，齐家兴，沈矗，等. 基于 MapReduce 和 Spark 的大数据主动学习比较研究[J]. 计算机工程与科学，2019(10)：44-45.
[4] 石焱，石宇强，夏世洪. 大数据背景下基于分布式 LDA 算法的生产模式识别[J]. 制造业自动化，2017(3)：24-28.
[5] 许力分，倪志伟，朱旭辉，等. 融合基于 MapReduce 并行改进二元蚁群算法与分形维数的属性选择方法[J]. 系统科学与数学，2019(6)：918-933.
[6] 张友海，李锋刚. 基于 MapReduce 的 KMeans 聚类算法的并行化实现[J]. 九江学院学报(自然科学版)，2017(1)：20-23.
[7] 刘炳含，付忠广，王永智，等. 基于并行计算的大数据挖掘技术及其在电站锅炉性能优化中的应用[J]. 动力工程学报，2018，38(6)：431-439.
[8] 肖文，胡娟，周晓峰. 基于 MapReduce 计算模型的并行关联规则挖掘算法研究综述[J]. 计算机应用研究，2018(1)：62-70.
[9] 向春梅，陈超. 基于 MapReduce 的改进 Eclat 算法[J]. 成都信息工程大学学报，2019(4)：369-374.
[10] 罗海洋. 基于 MapReduce 的关联规则挖掘算法的研究及应用[J]. 信息通信，2019(2)：30-33.
[11] 张占峰，耿珊珊. MapReduce 框架下常用聚类算法比较研究[J]. 河北省科学院学报，2019，36(2)：40-43.