

面向大数据的并行分类混合算法研究

陈学斌, 王 师, 董岩岩

(华北理工大学 理学院, 河北 唐山 063009)

摘 要: 针对传统分类算法及技术在处理海量异构数据存在的系统性能拓展性低、计算量大、耗时长、分类效果不佳等问题, 采用 Map-Reduce 与邻近分类算法融合设计适合大数据处理的并行分类混合算法, 利用加权欧氏距离并行计算, 达到提高海量数据分类效率、提高分类识别率和减小资源开销的目的, 搭建 Hadoop 集群研究并在多个数据集上测试算法的可行性. 实验结果表明, 并行分类混合算法在海量数据分类中显现出较好的分类效果, 是可行的海量数据分类模型.

关键词: 大数据; Map-Reduce; 算法融合; 并行分类

中图分类号: TP301.6

文献标识码: A

文章编号: 1000-7180(2016)04-0138-03

Research on Parallel Classification Hybrid Algorithm for Big Data

CHEN Xue-bin, WANG Shi, DONG Yan-yan

(College of Science, North China University of Science and Technology, Tangshan 063009, China)

Abstract: To solve the problem of the traditional classification algorithms and technologies in the huge amounts of heterogeneous data, such as low-expanding, large-calculating, time-consuming and poor classification. Parallel Classification Hybrid Algorithm is design by fusion of Map-Reduce and Nearest Neighbor Algorithm and parallel computation of weighted Euclidean distance, which improved the efficiency of mass data classification, improved the classification rate and reduced the cost of resource. The Hadoop platform is built for research and test the feasibility of the algorithm on multiple of data sets. The experimental results demonstrate that the Parallel Classification Hybrid Algorithm show a good classification effect in the massive data classification, is a feasible mass data classification model.

Key words: big data; Map-Reduce; fusion algorithm; parallel classification

1 引言

在移动互联网快速发展的带动下, 样式各异的数据急剧增长, 分类作为管理和分析数据的重要手段, 在商业、军事和科学研究等领域都有广泛的应用. 分类是对原有数据进行自动推导其内在描述, 并具有对后来数据预测归类功能. 通过建立模型描述数据集, 并利用建立的模型进行分类^[1]. 传统分类模型有贝叶斯、支持向量机、决策树等分类模型, 贝叶斯分类算法具有分类精度高、速度快、错误率低等

优势, 但分类准确率较低; 支持向量机分类模型具有分类准确率高、复杂度低的优点, 但是计算速度慢; 决策树模型能够很好地处理有噪声、复杂数据的优点, 但是处理大样本数据集时失效^[2-5]. 海量数据本身存在噪声、复杂、异构等特点, 传统方法因其算法局限、系统架构拓展性低、分类效果差的问题. 一方面在数据量大的情况下, 分类计算量大, 耗时长, 分类结果不尽如人意, 另一方面算法局限于同一类型数据, 拓展性差. 目前, 基于云计算的海量数据挖掘技术研究成为当下研究热点, Map-Reduce 是一种

高可扩展性的并行计算模型,可高效地处理海量数据,适合处理海量数据集,且能在可接受的时间内得出处理结果集^[6-7].

本文依托分布式计算的设计思想,解决海量数据分类系统架构拓展和分类效率的问题,利用邻近分类算法适合处理属性较多、数据量较大的分类问题的优点,引入 Map-Reduce 模型与其融合设计并行分类混合算法(Parallel Classification Hybrid Algorithm, PCHA). 解决传统算法在样本量少下建立模型预测后来数据分类的能力差问题,优化并提升原有分类算法的建模预测能力和分类识别率. 在超大数据集群下相对传统分类算法,利用 PCHA 算法可以提高并行分类运算效率,为海量数据分类提供并行计算框架模型,满足数据的下一步处理要求.

2 PCHA 算法提出

2.1 相关定义

定义 1 (加权欧氏距离) 设 $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$ 和 $x_j = (x_{j1}, x_{j2}, \dots, x_{jm})$ 为 n 维数据对象,若每一变量根据其重要程度分别赋予权重 $w_1, w_2, \dots, w_n$,则加权欧氏距离表示为

$$\text{dis}(x_i, x_j) = \sqrt{\sum_{k=1}^n w_k |x_{ik} - x_{jk}|^2}$$

定义 2(近邻度量) 设 n 维空间中有两个数据集对象 x_i 和 x_j ,则近邻度量表示为

$$\|x_i - x_j\| = \text{dis}(x_i, x_j)$$

式中,依据 $\|x_i - x_j\|$ 判定 x_i 和 x_j 是否相似.

定义 3(属性规范化) 设 A 为数值属性, $A_{\max}$ 和 $A_{\min}$ 分别表示属性最大值和最小值,为了防止大数据属性对分类的主导影响,根据属性规范化使得数值落入 $[0, 1]$ 区间中,属性规范化公式为

$$\tilde{A} = \frac{A - A_{\min}}{A_{\max} - A_{\min}}$$

2.2 算法设计

并行分类混合算法中的测试样本距离采用加权欧氏距离计算,结合 Map-Reduce 分而治之的设计思想,采取分散到 p 个并行计算节点上,训练样本总数为 n ,特征属性数为 s ,算法可进一步提高串行分类算法的计算效率,并有效地降低每一运算部分的时间复杂度和空间复杂度,即时间复杂度调整为 $O(ns)/p$ 和空间复杂度调整为 $O(n)/p$. 算法设计如下:

(1)Ready 阶段:本文提出的 PCHA 算法中利用加权欧式距离计算测试元组与训练元组的近邻度量

$$\|x_i - x_j\| = \sqrt{\sum_{k=1}^n w_k |x_{ik} - x_{jk}|^2}, \text{其中权重的选}$$

择视为数据对象 x_i, x_j 的重要程度而定,开始阶段采用所有数据对象视为同等的方式设置. 对属性数值进行属性规范化,做好数据准备工作.

(2)Input 阶段:实现 Map-Reduce 与邻近算法融合设计分类混合算法,搭建实现抽象类接口提供 Map 和 Reduce 两个函数,指明输入输出和其他运行参数. 通过 Input 阶段将大数据集划分为若干独立数据集,方便下一步处理,这里定义为 M 份数据集,并存储在 Hadoop 系统的 HDFS 中,提交给 JobTracker 后由相应 TaskTracker 执行任务.

(3)Map 阶段:在 Map-Reduce 框架中往往划分一个大数据集成小数据集,该阶段是对每组划分数据集的 $\langle \text{key}, \text{value} \rangle$ 对进行映射处理,TaskTracker 会调用空闲资源执行 Map 任务和 Reduce 任务. 这里的 Map 过程主要是对待分类数据集做键值对映射处理,即 $\langle \langle \text{节点}, \text{属性} \rangle, \langle \text{取值} \rangle \rangle$,任务根据各个属性规范化计算,且根据重要程度并行计算加权欧式距离,得到 $\langle \langle \text{节点}, \text{属性} \rangle, \langle \text{相似程度} \rangle \rangle$ 键值对,根据相似程度归类.

(4)Reduce 阶段:该阶段负责遍历所有 Map 阶段处理后产生的中间结果集,将根据同一 $\langle \text{节点}, \text{属性} \rangle$ 值排序、归并,统一输出 $\langle \langle \text{节点}, \text{属性} \rangle, \langle \text{相似程度} \rangle \rangle$,根据相似度来并行分类.

(5)Output 阶段:本阶段是与 Input 阶段呼应,即功能是对 Reduce 阶段输出结果集执行输出操作,输出存储到用户指定位置,此阶段得到的就是 PCHA 算法执行得到的分类的结果集,以便下一步处理.

3 实验结果及分析

3.1 环境配置

搭建 Hadoop 集群平台对提出的 PCHA 算法进行测试和验证算法的高效性和扩展性,测试算法在海量数据分类下的分类效率等,并与传统串行算法进行对比. 本文实验基于互联网舆情分析系统采集的海量数据作为分类原始数据集,采用了 1 GB、2 GB、4 GB 和 8 GB 等不同容量大小的数据集作为实验对象集. 选取部分作为实验训练集,部分作为测试集. Hadoop 集群节点配置如表 1 所示.

3.2 结果分析

本文提出的 PCHA 算法是为适应海量数据分类模型设计 Map-Reduce 与邻近算法融合设计并行

表 1 Hadoop 集群平台配置

Node	Capacity/ GB	Used/ GB	IP	State	Note
node1	186	2.43	10.0.0.1	normal	calculate
node2	186	1.01	10.0.0.2	normal	calculate
node3	186	4.11	10.0.0.3	normal	calculate
node4	186	83.15	10.0.0.4	normal	calculate
node5	186	4.72	10.0.0.5	normal	calculate
node6	186	2.38	10.0.0.6	normal	calculate
node7	186	1.37	10.0.0.7	normal	calculate
node8	186	8.83	10.0.0.8	normal	calculate
node9	186	34.02	10.0.0.9	normal	calculate
node10	186	28.85	10.0.0.10	normal	calculate
node11	186	5.12	10.0.0.11	normal	manage

分类混合算法,结合两者优势改进传统串行算法分类识别度低,计算量大,耗时长 的缺点,实验采用数据不同大小来测试,根据数据量的不同测试算法的可行性.通过 Hadoop 集群实现多节点并行运算,分类识别度对比统计结果如表 2 所示,从表 2 得知,传统串行算法识别度低,而 PCHA 算法识别度 优于串行算法.

表 2 算法识别正确率

次数	串行算法识别正确率/%	PCHA 算法识别正确率/%
1	75	80
2	70	79
3	75	81
4	73	80
5	76	82
6	76	81
7	72	79
8	77	81
9	75	82
10	76	83
平均值	74.5	80.8

从图 1 中可以看出,随着并行节点数的增加,运行时间呈递减趋势,相比串行算法而言,PCHA 算法运算速度快,耗时短,突显算法的优越性.但是由于 Hadoop 集群本身系统存在开销,面对少量数据

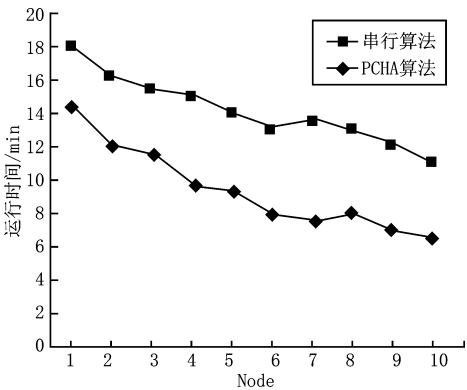


图 1 算法运行对比

时分类并行处理速度与串行运行速度相比,优势就不太明显,而面向海量数据并行处理时,效率明显优于串行算法,说明在海量数据分类下的并行算法处理大规模数据集的优势.

4 结束语

本文针对传统串行分类算法在处理海量数据存在的系统性能拓展性低、计算量大、耗时长、分类效果不佳等问题,研究 Map-Reduce 与邻近分类算法融合设计适应大数据处理的并行分类混合算法,介绍算法设计主要流程和方法等,搭建 Hadoop 集群研究并在多个数据集上测试算法的可行性.实验结果表明,PCHA 算法在海量数据分类中显现出较好的分类效果,相对于串行分类算法而言,具有较高的分类识别度和运行效率,是可行的海量数据分类模型.

本文需进一步研究方向:研究算法在跨节点运算上的并行化问题,从而进一步降低节点通讯耗时长;研究算法在数据样式各异间实现并行分类,从而进一步提高算法的扩展性问题.

参考文献:

[1] 陈刚,杨志强,刘秉权.一种基于 PLS 的概率神经网络分类算法[J]. 微电子学与计算机, 2015,32(5):73-78+83.

[2] 马宾,殷立峰.一种基于 Hadoop 平台的并行朴素贝叶斯网络舆情快速分类算法[J]. 现代图书情报技术, 2015,30(2):78-84.

[3] 李琼,陈利.一种改进的支持向量机文本分类方法[J]. 计算机技术与发展, 2015,25(5):78-82.

[4] 胡吉明,陈果.超球支持向量机文本分类方法改进[J]. 现代图书情报技术, 2014,30(9):74-80.

[5] 邹媛.基于决策树的数据挖掘算法的应用与研究[J]. 科学技术与工程, 2010,33(18):4510-4515.

[6] 李建江,崔健,王聘,等. MapReduce 并行编程模型研究综述[J]. 电子学报, 2011,39(11):2635-2642.

[7] 李玉丹,郑晓薇. Hadoop 下多模式并行分类算法及其应用研究[J]. 计算机工程, 2014,37(12):45-49.

作者简介:

陈学斌 男,(1970-),博士,教授.研究方向为大数据、云计算、物联网、网络安全等. E-mail:chxb@qq.com.

王 师 男,(1990-),硕士研究生.研究方向为云计算理论及应用.

董岩岩 男,(1990-),硕士研究生.研究方向为云计算理论及应用