面向云计算的数据挖掘分类算法研究*

卢 龙[1,2],王静宇[1],王 超[3]
(1. 内蒙古科技大学 信息工程学院,内蒙古 包头 014010;2. 中国北方稀土(集团)
高科技股份有限公司,内蒙古 包头 014010;3. 中国移动通信集团山东有限公司莱芜分公司,山东 莱芜 271100)

摘 要: 针对传统贝叶斯分类算法在处理海量数据时存在的运行时间长和分类准确率低等问题,在对传统的贝叶斯分类算法和云计算进行了深入研究后,提出了面向云计算环境的基于 MapReduce 模型的朴素贝叶斯分类算法。该算法实现了朴素贝叶斯分类算法的并行化,实现了大规模数据在云计算环境下的集群中进行贝叶斯分类处理。实验结果证明,该算法具有较高的分类准确率,在运行时间和加速比方面也有很好的效果。

关键词: 云计算; 朴素贝叶斯算法; MapReduce
中图分类号: TP393 **文献标识码:** A DOI: 10.19358/j.issn.1674-7720.2017.06.003
引用格式: 卢龙,王静宇,王超. 面向云计算的数据挖掘分类算法研究[J]. 微型机与应用,2017,36(6):7-9,12.

Classification algorithm of data mining based on Cloud computing

Lu Long[1,2], Wang Jingyu[1], Wang Chao[3]
(1. School of Information Engineering, Inner Mongolia University of Science &Technology, Baotou 014010, China;
2. China Northern Rare Earth (Group) High-Tech Co., Ltd., Baotou 014010, China;
3. Laiwu Branch, China Mobile Group Shandong Co., Ltd., Laiwu 271100, China)

Abstract: Aiming at the great challenge that comes from the large-scale automatic text classification of data mining classification algorithm because of the explosive increase of data, after studying traditional Bayesian classification algorithm, this paper proposed a Bayesian classification algorithm based on MaprRduce model for Cloud computing environment, making use of the Bayesian classification method of Cloud computing, dealing with the large-scale data in the Cloud cluster environment through Bayesian classification algorithm. Compared with single machine's handling time, experiments show that this method has greater execution efficiency which makes it suitable to process massive data rapidly in discretization way.

Key words: Cloud computing; naive Bayesian algorithm; MapReduce

0 引言

随着云计算、大数据等信息技术的快速发展,数据量呈现出了爆炸式的增长,数量级别从原来的 MB 级别迅猛增长到 TB 级别甚至是 PB 级别,这一严峻问题的出现给数据挖掘技术带来了前所未有的巨大挑战,海量数据的积累让人们有更多的数据可以利用,从这些海量数据中提取出对用户有价值的数据变得尤为重要。传统的数据挖掘算法通常要做的处理是先把数据从外存读入内存,然后进行分析处理,但是现如今数据量增大到惊人的级别时,由于对 CPU、内存等资源的急剧消耗,导致算法执行时间显著增加,算法的性能大幅度下降,根本无法达到用户的预期结果。在对海量数据进行挖掘处理时,要想获得理想结果,采用的数据挖掘算法必须要呈现出良好的可伸缩性和可并行性。云计算可以提供一种用于实现并行计算的模型[1],它将大规模数据的存储和计算能力均匀地分散到集群中,这些集群是由若干机器构成的,由许多的廉价机器搭建,在很大程度上降低了成本。云计算平台所具有的高速处理海量数据和计算海量数据这两大优势,更是为提高数据挖掘算法的效率和准确性提供了有力支撑,使传统的数据挖掘算法面临的难题得以解决。

数据分类作为一种重要的数据分析形式,常出现在数据挖掘领域中。目前比较常用的分类算法主要有:朴素贝叶斯分类算法、决策树分类算法、人工神经网络等。其中,在机器学习和数据挖掘研究领域,朴素贝叶斯分类算法是比较重要和常用的数据处理方法之一。朴素贝叶斯分类算法在简单、高效、分类效果稳定这个三个方面优势比较明显,它还具有牢固的理论基础,在实际应用中得到广泛的重视和应用。近年来,各国学者对贝叶斯分类方法展开了深入研究。

* 基金项目:国家自然科学基金项目(61662056,61462069);内蒙古自然科学基金(2015MS0622,2016MS0609)

文献[2]提出了一种基于K-means的贝叶斯分类算法，该算法主要思想是利用K-means聚类算法对原始数据集进行聚类分析，计算缺失数据子集中的每条记录与k个簇重心之间的相似度，把该记录分配到与其相似度最大的一个簇，用该簇中相应的属性均值来填充记录的缺失值，再用朴素贝叶斯分类算法对处理后的数据集进行分类，实验表明，在分类准确率上，与传统朴素贝叶斯分类算法相比，该算法分类准确率较高。文献[3]是基于Hadoop平台提出的朴素贝叶斯数据分类算法，该算法对特征选择方法进行了改进，并利用MapReduce编程模型实现了朴素贝叶斯并行分类算法。实验结果表明，该算法不仅提高了分类的正确率，而且在训练和测试集规模较大时体现出了很好的加速比，性能方面也有很大的提高。文献[4]则是通过对Hadoop基础平台MapReduce并行化编程模型进行深入研究后，对传统的朴素贝叶斯分类算法进行了MapReduce并行化改进，用以提高朴素贝叶斯分类算法对大规模数据处理的能力和计算效率。实验表明：改进后朴素贝叶斯分类算法在加速比和对中文网页进行分类识别率上都有很大的改进。

本文在对传统的贝叶斯分类算法和云计算相关技术深入研究后，提出了一种面向云计算并基于MapReduce模型的贝叶斯分类算法，利用提出的面向云计算的贝叶斯分类方法，对大规模数据在云计算环境下的集群中进行贝叶斯分类处理，通过实验证明该方法具有较高的执行效率。通过对大规模数据在云计算环境下的集群中进行贝叶斯分类处理，并对比大规模数据在不同节点上的运行时间和加速比可知，本文提出的算法具有较高的执行效率，平均分类正确率显著提高，适合用于海量数据的快速离散化处理。

1 相关工作

1.1 MapReduce编程模型

MapReduce是一种并行编程模型[5]，采用的是主(Master)/从(Slave)结构，在处理大规模数据时，将其分块后，分配到由普通机器组成的超大集群上并发执行。MapReduce编程模型主要分为两个阶段：Map和Reduce。Map阶段指的是映射，Reduce指的是规约；Map函数处理数据的形式是一个给定的键值对<key1,value1>，处理后生成另一个键值对<key2,value2>。随后MapReduce模型将Map阶段输出的相同的key2键值进行合并，形成一个新的键值对<key2,lsit(v2)>。Reduce函数则处理Map阶段合并的键值对<key2,lsit(v2)>，处理后形成一个形似<key3,value3>的键值对，并将这个键值对写入文件。

1.2 朴素贝叶斯数据分类算法

朴素贝叶斯算法(NBC)是简单常用的统计学分类算法[6]，该算法使用概率统计领域的知识对文本数据进行分类。它具体描述为：设样本数据集$D=\{D_1,D_2,\cdots,D_n\}$，对于某一特定$c\in C$，数据集有n个属性$A_1,A_2,\cdots,A_n$，测试样本$x=\{a_1,a_2,\cdots,a_n\}\in X$，则进行分类时，朴素贝叶斯分类器对每个类$c$计算后验概率：$p(x|c)$，为使$p(x|c)$估算有效，朴素贝叶斯分类算法通常假设数据各属性之间是相互独立的。则类条件概率$p(x|c)$表示为：

$$p(x|c)=\prod_{i=1}^{n}p(a_i|c)=p(a_1|c)\,p(a_2|c)\cdots p(a_n|c)$$

结合贝叶斯公式，朴素贝叶斯分类器对每个类c计算后验概率：

$$p(c|x)=\frac{p(c)\prod_{i=1}^{d}p(a_i|c)}{p(x)}$$

由于$p(x)$对所有的c是固定不变的，因此只要找出$p(c)\prod_{i=1}^{d}p(a_i|c)$最大类，这个类就是未分类$x$元组所属的类。

朴素贝叶斯分类算法包括两个过程：训练过程和测试过程[7]。训练过程是该算法消耗系统资源最集中的部分，如果将此过程带来的压力分解到一群机器上，那么在一定程度上会解决系统资源消耗严重的问题，而MapReduce并行编程模型完全可以实现此想法。

2 面向云计算的贝叶斯分类算法

随着信息技术的快速发展以及互联网的迅速普及流行，网络中的有价值数据越来越丰富，数据量也以指数的速度快速增长，这就给传统的数据挖掘算法带来了许多挑战和难题。为此需要通过对传统贝叶斯分类挖掘算法原理进行分析，寻找算法可并行的点，将传统的贝叶斯分类算法[7]应用到云计算环境下，利用MapReduce编程模式，实现传统贝叶斯分类算法的并行化，改变传统算法在单机模式上的局限，最终解决对海量数据训练和测试过程计算耗时长和分类正确率低的难题。

2.1 贝叶斯分类算法的可并行性

贝叶斯分类算法的训练过程是该算法消耗系统资源最集中的过程[8]，一般在处理一个或多个大文件时，顺序读写和计算会耗费很多系统资源，并且效率也不高。用户在处理大量数据时，数据文件间一般是独立的，可以借助数据分块的思想来处理。MapReduce并行编程模型完全可以实现此想法，这可以在很大程度上提高算法的运行速度和分类的准确率。

2.2 面向云计算的贝叶斯分类算法设计

利用Hadoop分布式系统基础架构中的HDFS存储能力和MapReduce的计算能力，实现对朴素贝叶斯分类算法的并行运行。贝叶斯的并行运行基本思路是[9]：由于训练数据集中数据之间并无关联，可以把数据集分块，这让传统的串行算法读取训练数据集的过程可以在多台机器上并行实现。在这个基础上，Hadoop分布式系统的处理机制对切块的数据分别处理。本文通过以下两部分实现朴

素贝叶斯分类算法的并行化，关键是将 MapReduce 模型引入其中：

第一部分： MapReduce 在处理大数据集时，Mapper 依照默认的输入格式读取大规模的文本数据，输入数据后再利用分词开源库对文本数据进行分词处理[10]，将近义词等去除，从而降低其特征维度，这个过程中还要建立相应的子目录，作为核心关键词的计算。为了实现筛选功能，在统计核心关键词时需要增加一个过滤器。按照 < word，class > 的形式输出，这一步的输出作为下一步操作的输入，Combiner 将一类的词归集在一起，计算出同一个 Mapper 的频率和，再把结果保存在 HDFS 中。过程如图 1 所示。

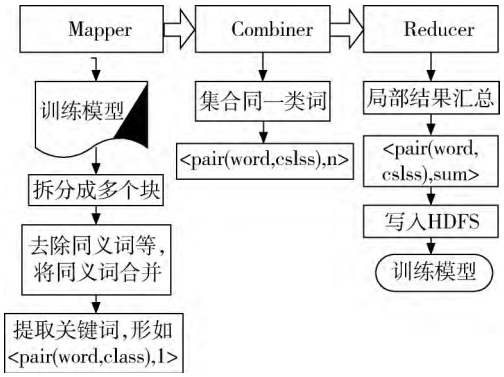

图 1 训练阶段

第二部分： 贝叶斯分类文本分类算法利用 MapReduce 模型进行并行实现，分类模型先由 Mapper 加载，然后将数据拆分成多个模块，分配给各个节点，通过模型向量计算出各个节点的概率和，以 < file，< Probability，class > > 的形式输出，作为 Mapper Task 的输入，其中 file 代表文件名，Probability 表示条件概率，class 代表类别。Combiner 将通过一个文件的特征集进行集中处理，完成 Reducer 任务，将全部条件概率计算出，得到先验概率后将数据保存在 HDFS。过程如图 2 所示。

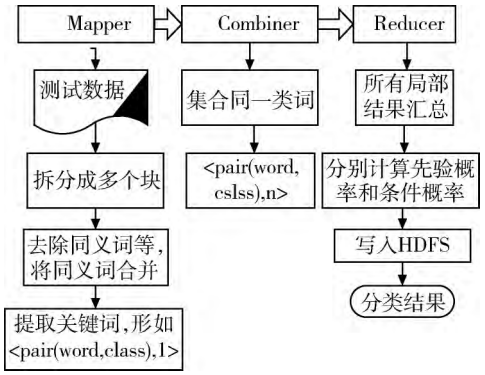

图 2 分类阶段

朴素贝叶斯分类算法并行实现后，算法中最耗时的运算被分布到集群中多个节点上，这在很大一定程度上减轻了一台机器的计算压力。假设训练集有 m 个类别，n 个特征项，待分类样本共有 k 个特征项，如果每个节点平均完成 t 个任务，则计算速度提高了 t 倍，复杂度为 $O(m^* n/k)$。

3 实验与结果分析

3.1 实验平台与实验数据

实验平台集群共 4 台计算机，搭建 4 个节点，软件环境为：操作系统：CentOS 6，JDK 版本：1.7，Hadoop 版本：2.5.2，Mahout 版本：0.10.1。搭建好 Hadoop 平台后开始运用 Mahout 实现本文提出的算法。Mahout 是 Hadoop 的一种高级应用[11]，运行 Mahout 需要提前安装好 Hadoop。Hadoop 安装完成，下面就开始运用 Mahout 运行本文提出的面向云计算的贝叶斯分类算法。实验数据来源于 UCI 中的 20-Newsgroups。

3.2 实验结果分析

本实验中，选取的文档集分别是 20-Newsgroups 中的 motorcycle，electronics，religion，baseball，politics。实验的分类准确性结果如表 1 所示。

表 1 并行朴素贝叶斯分类结果

	并行朴素贝叶斯分类结果		
类标签	测试文档数	正确分类数	正确率/%
motorcycle	202	185	91.58
electronics	178	166	93.26
religion	187	180	96.26
baseball	202	190	94.06
politics	187	172	91.98

从表 1 中可以得出，本实验的平均正确率为 93.54%，比传统的串行算法提高了 0.97%。

实验还对改进的贝叶斯分类算法分别从运行时间和加速比进行了测试。在 1 个、2 个、3 个、4 个节点上分别从运行时间和加速比方面对实验结果进行分析。从图 3 可以看出，随着节点数的增加，分类的运行时间显著缩短。同时，基于 MapReduce 模型的贝叶斯分类算法具有比较良好的加速比，实际加速比与理想的线性加速比十分接近，能够实现快速的分类，并且随着节点的增加，算法的加速比增长率逐渐变慢，这是因为 Hadoop 平台下节点之间的通信时间花费逐渐变大，同时实验选择的数据量越大，其算法的加速比增加越接近线性，如图 4 所示。

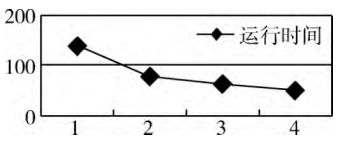

图 3 不同节点数下的运行时间

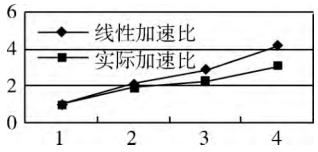

图 4 不同节点数下的加速比

4 结束语

传统的贝叶斯分类分类算法应用广泛，但其也存在运行时间长和分类准确率低等问题，云计算环境下，基于 MapReduce 模型的贝叶斯分类算法可以很好地解决此类

（下转第 12 页）

有的结果都是对相同的标签识别 10 次后取平均值的结果。

从图 3 可以看出，基于二分的 ALOHA 算法吞吐率最高稳定在 55% 左右，但是需要结合二分法，硬件要支持曼侧斯特编码，且算法的复杂度相对较高。其次是基于 EPC 的改进算法吞吐率稳定在 0.41% 左右，在工程实现中推荐这种方法，吞吐率较高且算法复杂度较低，对硬件要求也不高。如图 3 数学估计算法中标签估计最准确吞吐率能接近 35%，一般工程实现中不适用这种方法。

5 结论

本文首先介绍了 ALOHA 算法的基本原理和优缺点，然后介绍了 ALOHA 算法以及动态帧时隙的工作原理并对算法的理论吞吐率做了推导，得出当标签个数等于时隙数时 ALOHA 算法的理论最高吞吐率为 36.8%，简介了当前主流的基于 ALOHA 算法的改进算法原理，并总结了算法的优缺点，最后对不同吞吐率仿真得出基于二分法的 ALOHA 算法吞吐率最高接近 55%，但是实现难度较高，EPC 改进算法吞吐率其次，接近 41%，实现难度较低。本文可为 RFID 系统中防碰撞算法的研究工作提供参考。

参考文献

[1] 孙其博，刘杰，黎羲，等. 物联网：概念、架构与关键技术研究综述 [J]. 北京邮电大学学报，2010，33(3)：2-3.

[2] 黄玉兰，夏璞，夏岩，等. 物联网射频识别(RFID)核心技术详解 [M]. 北京：人民邮电出版社，2012.

[3] CHA J R，KIM J H. Dynamic framed slotted ALOHA algorithm using fast tag estimation method for RFID system [C]. Consumer Communications and Networking Conference，2006 CCNC. 3 IEEE，2006：768-772.

[4] 陈平华，王康顺，李超，等. 基于线性回归的动态帧时隙 ALOHA 算法 [J]. 计算机仿真，2014，31(7)：259-263.

[5] 石封查，崔琛，余剑. 基于标签运动的一种新型 RFID 防碰撞算法 [J]. 计算机科学，2013，40(6)：76-79.

[6] EPC Global. EPC Radio-Frequency Identity Protocols Genger-nation-2 UHF RFID Protocol for Communication at 860 MHz ~ 960 MHz Version [S]. California：EPC Global，2008.

[7] EPC Global. EPC Radio-Frequency Identity Protocols Genger-nation-1 [S]. American：UCC，2013.

[8] 吴淼，刘德盟，张钊锋. 基于 EPC Gen2 防碰撞机制的研究与优化 [J]. 微电子学与计算机，2013，30(5)：101-104.

[9] 徐圆圆，曾隽芳，刘禹. 基于 Aloha 算法的帧长及分组数改进研究 [J]. 计算机应用，2008，28(3)：588-590.

[10] 鞠伟成，俞承芳. 一种基于动态二进制的 RFID 抗冲突算法 [J]. 复旦大学学报(自然科学版)，2005，44(1)：46-50.

(收稿日期：2016-09-27)

作者简介：

杨晓娇(1988 -)，通信作者，女，硕士，助理工程师，主要研究方向：无线传感网、RFID。E-mail：yangxiaojiao@cqjtu.edu.cn。

吴必造(1985 -)，男，硕士，软件工程师，主要研究方向：物联网安全、物联网传输、RFID。

(上接第 9 页)

问题，缩短数据处理时间和提高数据处理的分类准确率。实验结果表明，随着节点的增加，并行化的贝叶斯分类算法的数据处理的运行时间在不断下降，实际加速比与理想的线性加速比十分接近，该算法具有较高的执行效率，平均分类正确率为 93.54%，比传统的串行算法提高了 0.97%。因此，并行化的贝叶斯分类算法在处理海量数据时具有现实意义。

参考文献

[1] 冯登国，张敏，张妍，等. 云计算安全研究 [J]. 软件学报，2011，22(1)：71-83.

[2] 张亚萍，胡学钢. 基于 K-means 的朴素贝叶斯分类算法的研究 [J]. 计算机技术与发展，2007，17(11)：33-35.

[3] 张红蕊，张永，于静雯. 云计算环境下基于朴素贝叶斯的数据分类 [J]. 计算机应用与软件，2015，32(3)：27-30.

[4] 江小平，李成华，向文，等. 云计算环境下朴素贝叶斯文本分类算法的实现 [J]. 计算机应用，2011，31(9)：2551-2554.

[5] DEAN J，GHEMAWAT S. MapReduce：simplified data processing on large clusters [C]. Proceedings of Operating Systems Design and Implementation (OSDI)，2004：107-113.

[6] 郑炜，沈文，张英鹏. 基于改进朴素贝叶斯算法的垃圾邮件过滤器的研究 [J]. 西北工业大学学报，2010，28(4)：622-627.

[7] 张增伟，吴萍. 基于朴素贝叶斯算法的改进遗传算法分类研究 [J]. 计算机工程与设计，2012，33(2)：750-753.

[8] 李军华. 云计算及若干数据挖掘算法的 MapReduce 化研究 [D]. 成都：电子科技大学，2010.

[9] 蒋婉婷，孙蕾，钱江. 基于 Hadoop 的朴素贝叶斯算法在中文微博情感分类中的研究与应用 [J]. 计算机应用与软件，2015，32(7)：60-62.

[10] 向小军，高阳，商琳，等. 基于 Hadoop 平台的海量文本分类的并行化 [J]. 计算机科学，2011，38(10)：184-188.

[11] RUI M E，RUI P，RONG C. K-means clustering in the Cloud-A Mahout test [C]. IEEE Workshops of International Conference on Advanced Information networking and Applications，IEEE，2011：514-519.

(收稿日期：2016-09-26)

作者简介：

卢龙(1982 -)，通信作者，男，硕士研究生，主要研究方向：云计算与数据挖掘。E-mail：btu_wjy@qq.com。

王静宇(1976 -)，男，博士研究生，副教授，主要研究方向：云计算与信息安全。

王超(1987 -)，男，硕士研究生，助理工程师，主要研究方向：云计算与数据挖掘。