云计算环境下海量数据挖掘分类算法研究

高文强　张晓梅

（宿州学院信息工程学院，安徽 宿州　234000）

摘　要：目前互联网数据的不断膨胀已成为趋势，如何从海量数据中挖掘有效信息成为研究热点。笔者旨在通过 MapReduce 框架在云计算基于开源 Hadoop 平台下能够有效而且快速地工作，保证算法的高效执行，挖掘有效数据。

关键词：云计算；分类算法；MapReduce；Hadoop

中图分类号：TP311.13　　**文献标识码：**A　　**文章编号：**1003-9767（2016）15-096-02

1　概述

随着云计算的产生，互联网数据存储形式越来越多样，存储容量也日益庞大，如何从海量的数据中挖掘出有效数据，采用何种算法直接关系到挖掘过程的效率和精确度。

本文旨在在云计算环境下设计了 MapReduce 化模型框架，采用朴素贝叶斯算法以 Java 思想的 Strategy 设计模式，并根据该分类模块对朴素贝叶斯算法进行实验测试，以充分提高挖掘网络数据的高效性和准确性。

2　相关理论研究

2.1　云计算 Hadoop 平台

云计算（Cloud Computing）是基于互联网技术的崭新的服务模式，是分布式计算、并行计算、网络存储、虚拟化等传统技术相互融合的产物。

Hadoop 是一个能对海量数据进行分布式处理的软件框架，对数据处理可靠、高效、可伸缩。Hadoop 核心部分由 Hadoop Common、MapReduce、HBase、HDFS 和 ZooKeeper 组成，能够实现分布式数据存储、分布式计算模型和并发控制。Hadoop 基本框架如图 1 所示：

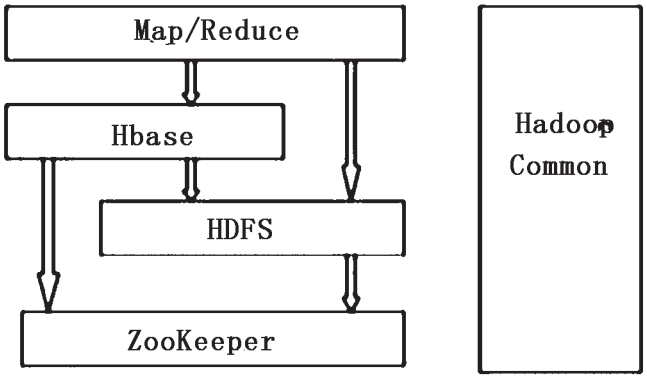

图 1　Hadoop 基本框架

2.2　数据挖掘

数据挖掘是人工智能领域知识发现过程的一个应用，通过分析每个数据，从海量数据中寻找其规律，发掘有价值的信息并据此制定相应的对策。

2.3　MapReduce 模型

Hadoop 是知名的云计算开源系统平台，Map/Reduce 算法是其关键技术，可以抽象为 Map 函数和 Reduce 函数。Map 由使用者编写，是一个分解过程，将一个较大的文件分解成若干个子文件，由计算机进行分布式处理。再通过 Reduce 函数将所有的子文件数据汇总，并输出最终结果。其工作原理如图 2 所示：

3　MapReduce 模型上分类算法的执行

3.1　MapReduce 模型上分类算法的执行框架

MapReduce 模型上分类算法的处理过程主要是由划分阶段、Map/ 构建基本分类器阶段、Reduce/ 集成阶段三部分组成。在划分阶段，根据置换的抽样方式，把数据集 D 划分成 m 个子集 {D1,D2,⋯,Dm}，即 D={D1,D2,⋯,Dm}，m 值由用户指定；在 Map/ 构建基本分类器阶段，采用分类算法，每一个 Map 任务在数据集 Di 上构建一个基本的分类器 Ci(i ∈ [1,m])；在 Reduce/ 集成阶段，最终把 m 个基本分类器集合成一个统一的分类器 C。

3.2　朴素贝叶斯算法

本文采用的是朴素贝叶斯分类算法，实现该算法需要 Map 接口函数和 Reduce 接口函数。

（1）Map 接口函数：首先输入训练数据集，并给训练数据集标注类别标签或者标签、属性名和属性值的组合 key，标注标签出现的频率 value。通过解析标签构建一个标签、

基金项目：省级创业创新项目“云计算环境下海量数据挖掘分类算法研究”（项目编号：201510379096）。

作者简介：高文强（1994-），男，安徽亳州人，本科在读。研究方向：数据挖掘。

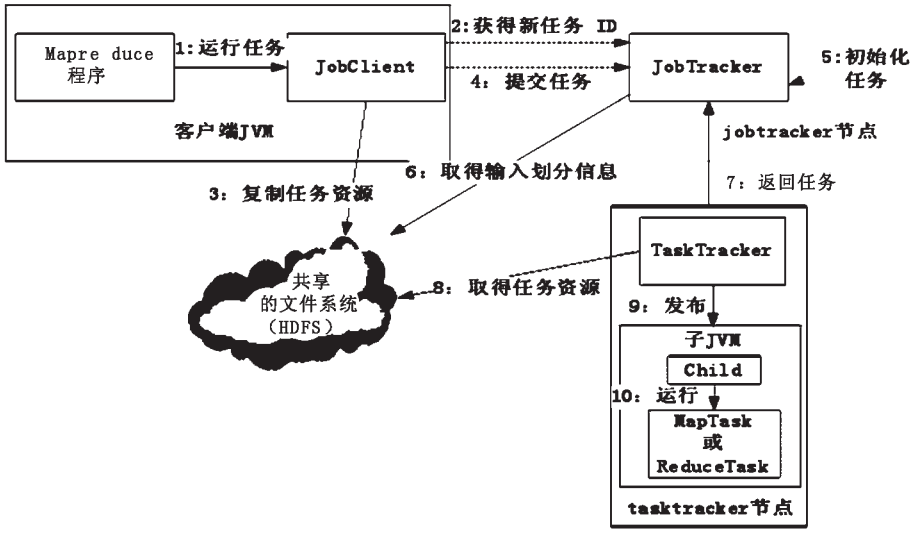

图 2 Hadoop 平台上 MapReduce 的工作原理

属性名和属性值的一个连接字符串 key，频率 value 设置为 1，输出 <key,value> 键值对。

（2）Reduce 接口函数：首先输入由 Map 接口函数分别输出的 key 和 value，设置一个计算器 sum 初始化为 0，记录 key 的当前统计频率，通过循环训练，把 key 赋值给 key ’，sum 赋值给 value ’，输出 <key ’,value ’> 键值对，其中 key ’ 标志标签或标签、属性名和属性值的组合，value ’ 标志统计频率结果。

3.3 性能测试

在 Hadoop 平台上，基于 MapReduce 模型进行测试，实验采用线性加速比作为衡量标准，实验目的是验证采用并行化的算法在云计算平台上提高执行效率和性能。结果显示出基于 MapReduce 的朴素贝叶斯算法在训练过程中相关的加速比趋向于线性加速比，分类效果较好，大大提高了执行效率，如图 3 所示。

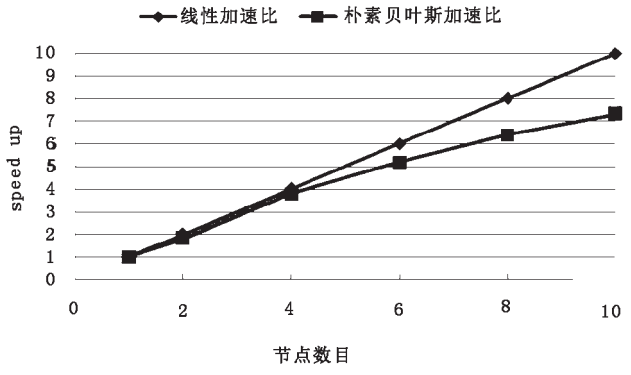

图 3 性能测试结果

4 结 语

随着云计算大数据时代的到来，信息的高速膨胀无疑成为了一把双刃剑，一方面带给人们众多的知识，另一方面又存在大量无用的信息也直接影响着人们的生活，如何从海量的数据中提取有用信息至关重要。传统的数据挖掘方法对已有数据进行统计分析，辅助人们制定决策，但是对于海量数据处理存在不适应问题。

分类算法作为数据挖掘中极其重要的一部分，在云计算环境下，借助 Hadoop 平台的 MapReduce 模型，通过在该平台上对所研究的分类算法进行 MapReduce 处理，利用朴素贝叶斯算法中的 Map 接口函数和 Reduce 接口函数进行模拟训练，分类效果较好，在一定程度上提高了数据利用效率。但是算法在执行过程中需要对原始数据进行分割，然后分配给各个处理器处理，处理结束过后又进行统一收集处理，在这个过程中可能会导致数据分类的精确度降低，因此，这是今后努力的方向。

参考文献

[1] 杨善林 , 罗贺 , 丁帅 . 基于云计算的多源信息服务系统研究综述 [J]. 管理科学学报 ,2015(5).

[2] 薛玉 . 云计算环境下的资源调度优化模型研究 [J]. 计算机仿真 ,2013(5).

[3] 申丽君 , 刘丽 , 陆锐 , 等 . 基于改进免疫进化算法的云计算任务调度 [J]. 计算机工程 ,2012(9).

[4] 吕良干 . 云计算环境下资源负载均衡调度算法研究 [D]. 乌鲁木齐 : 新疆大学 ,2010.

[5] 田冠华 , 孟丹 . 詹剑锋 . 云计算环境下基于失效规则的资源动态提供策略 [J]. 计算机学报 ,2010(10).