文章编号:1008-0570(2012)09-0486-02

基于 Canopy 的 K-means 多核算法

Canopy for K-Means on Multi-core

(浙江传媒学院) 邱荣太
QIU Rong-tai

摘要: 基于 Map-reduce,提出了面向多核处理器应用于大规模集群的并行编程方法,应用该方法运行数据挖掘算法 Canopy 和 K-means。针对 K-means 算法对初始聚类中心敏感,提出了基于 Canopy 的 K-means 优化算法。基于实际数据集的实验结果表明,多核 Canopy-K-means 聚类算法的准确度和执行效率随着核数的增多呈线性增长。

关键词: K-means; Map-reduce; 多核; Canopy
中图分类号: TP309 **文献标识码:** A

Abstract: In this paper, we develop a applicable parallel programming method which based on Map-reduce, one that is easily applied to machine learn algorithms Canopy and K-means on multi-core and large cluster. A improved K-means algorithm based on Canopy is presented according to it's sensitiveity to the initial centers. Our experimental results show basically linear speedup with an increasing number of processors.

Key words: K-means; Map-reduce; Multi-core; Canopy

技术创新

1 引言

随着多核处理器的广泛应用,计算机性能的提高将更多地依赖于处理器的数量的增加.本文研究在多处理器机群上如何并行编程。核心目标是在聚类算法方面应用 Map-reduce 编程框架以便高效的利用数量众多的可并行的处理器核,允许程序员轻松高效地开发一个多核机群机器学习算法,而不用编写专门的并行任务执行优化。本文创新如下:

(1) 传统的 K-means 算法对初始聚类中心敏感,针对 K-means 算法存在的问题,利用 Canopy 聚类划分来优化初始聚类中心。

(2) 由于先将所有的数据点进行 Canopies 有覆盖划分,在计算数据点离哪个 K-center 最近时,不必计算其到所有 K-centers 的距离,只计算和它在同一个 Canopy 下的 K-centers 距离,其效率也大大提高。

(3)存储的数据各部分并不相互依存,可以使用 Map-reduce 编程模型高效地进行并行处理。

(4)随着处理器核数及处理节点的增多,我们得到的执行效率符合线性增长。

2 基于 Mapreduce 的 CanopyKmeans 算法

2.1 算法思想

首先所有数据集构成一个数据空间,将这个数据空间按某种特征划分成 K 类,每一类的数据成员具有一些共同特征。初始数据根据数据存储节点及 mapper 核的个数 P 划分成 $C_1 C_2 \ldots C_p$ 个数据集,每个数据集可以分别分配给某一 Mapper 节点独立执行。选择一个节点作为 master 节点,负责并行任务的调度工作,在 master 节点处设置两个共享文件,这两个文件可以被所有处理器访问,其中一个文件包含初始随机选的 K 个聚类中心点和每次迭代执行后产生的 K 个聚类的聚类中心。这是一个全局文档,每次迭代都在这个文件后面增加一些记录。另一个文件保存产生 Canopies 列表,这些信息提供全局调度作用。Master 节点负责管理各个 Mapper 节点和 Reduce 节点的调度及数据分配与结果的回收,分配给 Mapper 核相应的小数据块,并且传给每个 Mapper 节点 K 个聚类中心以及 Canopies 列表。每个 Mapper 核并行计算各自要求的任务后通知 Master,然后 Master 调度相应的 Reduce 进行处理,返回结果以决定是否下一轮迭代。

首先用 Canopy 聚类技术分两个阶段执行聚类:第一步就是粗糙地、快速和近似地把数据分成一些重叠的子集,称之为罩盖 (Canopy);然后对 Canopy 内的点用精确的计算方法再聚类。它在两个阶段使用了不同的距离度量方法,形成重叠的 Canopy.创建好 Canopy 后,第二步就是针对 Canopy 内的点使用 K-means 算法进行聚类。这样只需要对罩盖内的点进行精确的计算,从而大大减少了传统聚类算法中对所有数据点进行精确计算的计算量,另外允许有重叠的子集也增加了算法的容错性和消除孤立点作用。从某种意义上说,只要保证第一步的准确性,整个算法就是准确的,通过实验表明,使用这种罩盖技术后,不仅极大地减少了距离的计算量,其准确程度比一般的聚类方法还略有提高。现在我们用 Map-reduce 模型来执行 Canopy 和 K-means 集群算法.

2.2 算法流程

2.2.1 数据预处理

将数据空间的数据进行预处理,每一个数据写成特定的格式(数据,出现次数)即(datavalue,emergetime),将每个数据按 key 为 datavalue 进行聚合,聚合后的每个数据为数据和出现次数的组合。

Mapper:(offset,datavalue)-->(datavalue,1)

邱荣太:工程师 硕士

Reducer: (datavalue,1) --> (datavalue,emergetime)

2.2.2 产生 Canopies

(1) 总体思路:预处理后的数据,每一个数据值(即 datavalue)可以看成一个中心点,按 Canopy 算法产生一些可覆盖的划分,初始划分以随机选的第一个数据作为本聚类标识 Canopyid。这样对于每个 Mapper 节点的数据,都要判断它是否落入之前产生的 Canopies 中,若落入其中任意一个 Canopy 则标识自己相应的 Canopyic,否则产生一个新的 Canopy,并用自己的数据值作为 Canopid。

(2)实现过程:在 Master 计算节点上存储已产生的 Canopies,这样若某个节点产生新的 Canopid 要及时传送给 Master 节点,以便其它节点读取。

(3)实现方法:

Mapper: (datavalue,emergetime)-->(Canopyid,(datavalue,emergetime))

Reducer: (Canopyid,(datavalue,emergetime)) -->

(Canopyid,(datavalue,emergetime),…, (datavalue,emergetime))

2.2.3 将数据分配到 Canopies

(1) 总体思路:经过 2.2.3.2 处理后的数据,整个数据空间的数据划分成了几个 Canopies,这些 Canopies 存储于 Master 中。每个 Mapper 节点读取 Master 中的 Canopies,然后判断每一个数据落在哪几个 Canopies 上,并输出数据和相应的 Canopies (用其 Canopid 表示)。

(2)实现过程:Mapper 计算节点上每个数据所属的 Canopid。

(3)实现方法:

Mapper: (datavalue,emergetime) -->(datavalue, (emergetime, Canopyid_list))

Reducer: (datavalue,(emergetime, Canopyid_list)) -->

(datavalue,(emergetime, Canopyid_list))

2.2.4 实现 Canopy-Kmeans 聚类

(1) 总体思路: 根据 Canopy 算法产生的 Canopies 代替初始的 K 个聚类中心点,由于已经将所有数据点进行 Canopies 有覆盖划分,在计算数据离哪个 k-center 最近时,不必计算其到所有 k-centers 的距离,只计算和它在同一个 Canopy 下的 k-centers。这样可以提高效率。

(2) 实现过程:

每次 Mapper 节点在进行 map()前需要通过 Master 节点加载上一次产生的 K-centers 和产生的 Canopies 列表,并计算每个 K-center 落入哪些 Canopies 中。

每次迭代过程 mapper 的输入数据都是 2.2.3.3 产生的 (datavalue,(emergetime, Canopyid_list))。每个数据只用与落入该数据在同一个 Canopy 中的 K-centers 比较距离,选择最短的那个 K-center 作为要划分的类,再以 K-center 作为 key,产生一条格式为:"k-centerid dativalule emergetime Canopyid_list"的记录。

距离试题函数:每个数据点的 dativalule 之间差的绝对值。

(3)实现方法:

Mapper: (datavalue,(emergetime, Canopyid_list))-->

(k-centerid,(datavalue,emergetime, Canopyid_list))

Reducer: (k -centerid, (datavalue,emergetime, Canopyid_list)) -->

(new_k-centerid,(datavalue,emergetime, Canopyid_list))

(3)最终结果的输出:

我们可以把每次迭代后产生的 new_k-centers 与上一次的 k-centers 作比较,当他们的距离濒于预先设定的阈值时,认为迭代结束,输出最后一次的 (k -centerid, (datavalue,emergetime, Canopyid_list))格式为:"k-centers,datavalue_list"。

2.3 算法复杂性分析

K-Means 法使用随机方式选择 K 个数据作为初始的聚类中心,按照算法的迭代执行,整个算法的结束条件是类的重心(或凝聚点)不再改变。传统的 K-Means 的计算复杂性是 O(ukt),其中,u 为文献数量,k 为类的数量,t 为迭代次数。在运用 Canopy 算法对 K-Means 进行优化的情况下,由于在划分 Canopies 是可覆盖划分,即某一点有可能同时属于 n 个 Canopies,这样聚类需要比较 ukn2t/c 次,其中 C 为 Canopies 的数量. 在多核框架下,其理论复杂性是 $O(ukn^2t/cp+ukn^2\log(p)/c)$,其中 P 为处理器核的数量。K-Means 在多核并行执行情况下其效率将近是单核下的 P 倍,随着核数的增多,效率得到线性增长。

3 实验与分析

为了方便比较,我们把算法分两种情况下执行:一个在我们设计的框架下,另一个以传统的方式顺序执行。实验数据从网络上采集,在实验开始之前在实验环境的存储设备上准备完毕。网络上的数据主要是两个部分,一部分是 Wikipedia 的数据,一部分是学校学生成绩的数据。Wikipedia 的数据大概有 10G 多,存放在 Hadoop 的分布式文件系统 HDFS 中。学校学生成绩数据也有将近 800M,同样放在 HDFS 中。我们进行了大量的试验来测试其效率。不仅如此,我们还分别在拥有不同核个数的处理器上执行。我们在 Intel X86 ,Intel (R) Core™ Duo CPU 2 MHz CPUs , 2GB 物理内存 。并且在运行 HP-UNIX 的机器 HP9000 上分别做实验。另外我们通过 4,6,8,16 核的处理器上模拟试验得到如下结果。在双处理器执行过程中,执行各个数据的平均效率比原来大约提高了 1.9 倍。这是由于原始的 K-means 算法并没有很好的利用处理器时钟周期,然而在我们把任务独立分给不同处理器并行的情况下算法效率得到了很大的提高。通过在 2,4,8,16 多处理器机器上实验,我们可以很直观地把我们得到的结果描绘成图 2.在图中,实线代表平均加速倍数,虚线分别代表最大与最小的加速倍数。分析图中可知,随着处理器个数的增加,我们得到的加速倍数是线性增长的,然而这个加速斜率略低于 1。通过分析可知,没有得到完整的线性增速是由于处理器之间相互通信和集合阶段没有并行的原因。在多核的情况下我们通过模拟试验,结果比现在更理想,因为他们之间共享存储区域,其相互通信比以前少。

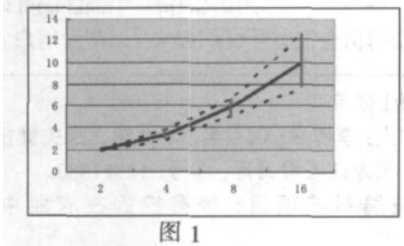

图 1

4 结语

即使处理器的频率没有提高,同样可以提高机器学习算法的效率而不受限于处理器的频率。 (下转第 233 页)

技术创新

NPP 的高值区,主要分布在延庆东北部、房山西部、怀柔、昌平北部及西部地区、怀柔、门头沟地区,这些地区海拔相对较高,属于山区,林地,即灌丛、阔叶林、针叶林、混交林等多分布于此,故 NPP 相对较高。而朝阳区、海淀区、丰台区等繁华地带,NPP 较低,这些区域主要是商业用地和居民用地,植被覆盖少。因为这些地区地势平坦,耕地大都分布于此,但耕地作物季节性较强,故全年尺度上 NPP 积累不多。

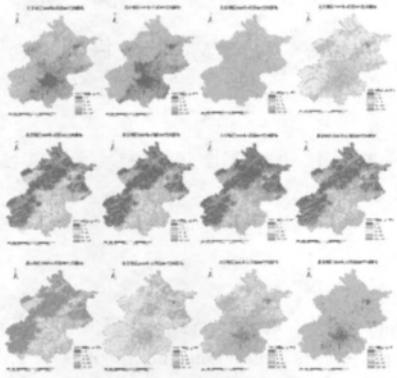

图 5 北京市 2009 年 1~12 月份 NPP 空间分布(单位:gC/m²·月)

5 结论

本文以北京地区为例,采用 HJ-1/CCD 影像数据,并基于其空间分辨率与成图比例尺之间的关系, 对 CASA 模型中最大光能利用进行修订,反演 2009 年 12 个月份耕地、林地、草地、水域、居民用地、其他未利用土地六种地物类型的 NPP,并从时空分布上进行分析。实践证明,采用面积加权平均方法修订最大光能利用率可提高 NPP 估算精度, 且 HJ-1/CCD 高时空分辨率特点,可广泛应用于 NPP 大尺度、高时间分辨率遥感反演。

本文创新点:本文基于 HJ-1 影像数据,针对北京地区不同地物类型在 CASA 模型中重要参数最大光能利用率进行修订。

作者声明:"作者对本文版权全权负责,无抄袭。"

参考文献:

[1] 柯金虎,朴世龙,方静云.长江流域植被净第一性生产力及其时空格局研究[J]. 植物生态学报, 2003,27(6)764~770.
[2] 李世华,牛铮,李壁成.NPP 过程模型遥感驱动因子分析[J].水土保持研究,2005 年 6 月第 12 卷,第 3 期.
[3] 朱文泉,潘耀忠,龙中华等. 基于 GIS 和 RS 的区域陆地植被 NPP 估算——以中国内蒙古为例[J]. 遥感学报,2005,5.
[4] 朱文泉,潘耀忠,张锦水. 中国陆地植被净初级生产力遥感估算[J]. 植物生态学报,2007,31(3)413~424.
[5] 孙睿, 朱启疆. 气候变化对中国陆地植被第一性生产力影响的初步研究[J]. 遥感学报,2001.1(5).
[6] 王桥,张峰等. 环境减灾-1A、1B 卫星环境遥感业务运行研究[J].航天器工程,2009,6:125-132

作者简介: 米晓飞(1985-),女,硕士研究生,研究方向:生态遥感。

Biography: MI Xiao-fei (1985-), female, Research direction is Ecological remote sensing Remote Sensing.

(100101 北京 中国科学院遥感应用研究所) 米晓飞 余 涛 李慧芳 李家国 张树凡

(454000 焦作 河南理工大学测绘学院) 李慧芳 袁占良

通讯地址:(100101 北京市朝阳区大屯路天地科学园区 遥感应用研究所 B 座 402 室) 米晓飞

(收稿日期:2011.09.28)(修稿日期:2011.12.28)

(上接第 481 页)

[2]段海滨.蚁群算法原理及其应用[M].北京:科学出版,2005,12.
[3]徐红梅,陈义保,刘加光,王燕涛.蚁群算法中参数设置的研究[J]. 山东理工大学学报(自然科学版),2008,1
[4]柳长源,毕晓君,韦琦.基于蚁群算法求解 TSP 问题的参数优化与仿真[J].信息技术,2009,4
[5]王霄,吴开军等.蚁群算法及其在旅行商问题(TSP)中的应用(城市个数,70)[J].微计算机信息,2010,11-3:199-201

作者简介: 郝春梅(1975-),女(汉族),黑龙江人,哈尔滨金融学院计算机系副教授,硕士,主要研究领域为管理信息系统、电子商务 ;吴 波(1972-),男(汉族),黑龙江人,哈尔滨金融学院计算机系副教授,硕士,主要研究领域为电子商务。

Biography: HAO Chun-mei(1975-),Female(the Han nationality),Heilongjiang, Computer Science Department of Harbin Finance University , Associate Professor,Master, Research area: management information service,E-commerce.

(150030 哈尔滨 哈尔滨金融学院计算机系)郝春梅 吴 波

通讯地址:(150030 黑龙江省 哈尔滨市 香坊区电碳路 65 号哈尔滨金融学院计算机系) 郝春梅

(收稿日期:2011.09.28)(修稿日期:2011.12.28)

(上接第 487 页)

通过多核处理器可以计算更大的数据集和获得更好的效率。利用谷歌的 Map-reduce 编程方法,使得 K-means 算法在双核的情况下效率提高了两倍, 在 64 核的情况下得到 56 倍的执行效率,得到线性增长的效率,并且对的初始点用 Canopy 算法进行优化提高了聚类的准确度。算法具有很大的适用范围,本文创新之处是我们并没有对原来的算法没作根本的改变, 只是提出应用 Map-reduce 并行执行 K-means 聚类算法以及对 K-means 初始聚类中心优化方法,显著提高了执行效率。

本文无抄袭,作者全权负责版权事宜。

参考文献:

[1] Jeffrey Dean, Sanjay Ghemanwat, MapReduce: Simplified Data Processing on Large Clusters
[2] 张建民. 一种改进的 K-means 聚类算法 [J]. 微计算机信息,2010,3-3,2.
[3]Kenneth Heafield Hadoop Design and K-Means Clustering Google Inc January 15 2008
[4] Bradley, Fayyad, Refining Initial Points for K-Means Clustering 1998.5
[5] Dummler, Rauber, Runger, Mapping Algorithms for Multiprocessor Tasks on Multi-core Clusters
[6] 丁光华,周继鹏,周敏.基于 MapReduce 的并行贝叶斯分类算法的设计与实现[J]. 微计算机信息 2010,3-3,3

作者简介: 邱荣太(1982-),男(汉族),浙江传媒学院播音学院实验办,工程师,硕士,主要研究分布式及流媒体应用。

Biography: QIU Rong-tai(1982-),ZheJiang, Zhejiang University of Media and communications),distributed network and p2p stream media.

(310018 浙江 杭州 浙江传媒学院) 邱荣太

通讯地址:(310018 杭州市下沙高教园区学源街 998 号播音主持艺术学院) 邱荣太

(收稿日期:2011.09.28)(修稿日期:2011.12.28)

技术创新