

文章编号:1673-0062(2017)01-0081-05

基于支持向量机的任务调度模型

牟式标,金伟健

(义乌工商职业技术学院 机电信息学院,浙江 义乌 322000)

摘要:云计算框架大大改进了并行算法的实现难度,但是大部分算法有其局限性.介绍了 MapReduce(映射化简)的基本实现原理和调度模型的缺陷,提出了基于支持向量机的 MapReduce 进化算法,并给出了基本模型及实现.运用 Hadoop 云计算平台进行了仿真验证,实验结果表明,基于支持向量机的 MapReduce 计算框架在候选云节点的调度分配的准确性上有明显提高,并且加快了数据迭代的效率.

关键词:云计算;MapReduce;支持向量机;Hadoop

中图分类号:TP302.1 **文献标志码:**A

The Tasks Scheduling Model Based on Support Vector Machine

MU Shi-biao, JIN Wei-jian

(School of Electro-mechanical and Information Technology, Yiwu Industrial and Commercial College, Yiwu, Zhejiang 322000, China)

Abstract: The cloud framework reduced the difficulty to realize parallel algorithm. But most of the algorithms have the defects. The fundamental and scheduling model of MapReduce are introduced. The evolving algorithm is proposed based on Support Vector Machine. The basis model and realization are built. By simulation realization on Hadoop cloud computing platform, compared with the traditional scheduling algorithm, experimental results show the new theory based on the SVM improve the accuracy of cloud candidate point scheduling and improve the speed of data iteration.

key words: cloud computing; MapReduce; SVM; Hadoop

云计算是当前 IT 领域最热门的话题之一,代表了 IT 技术发展的方向,它最重要的理论就是“网络即计算机”,将大规模的易用的可访问的虚

拟资源连接起来形成一个资源池(包括计算资源、存储资源、软件资源等),为终端计算机提供“召之即来,挥之即去”且似乎“能力无限”的 IT

收稿日期:2016-09-15

基金项目:浙江省 2016 年度高校国内访问工程师“校企合作项目”(FG2016128)

作者简介:牟式标(1979-),男,讲师,硕士,主要从事计算机应用的研究.E-mail:164297164@qq.com

服务^[1].

虽然云计算被大家大力推广到各个前沿科技,云计算也确实能为社会创造巨大的价值,但是云计算的发展也面临很多问题,例如:云节点的任务调度问题.因为只要这个云计算的资源池越大,被终端用户依赖程度越高,那么就需要考虑访问该资源池的效率问题.云节点的分配,网络节点处理请求的能力,选择怎样的空闲云节点更加高效等问题就值得关注.

文章提出了一种基于支持向量机的 MapReduce 任务调度模型,其基本思想是对云节点进行特征值建模,同时考察多维节点的情况,动态分配核函数值高的资源池,拒绝可信度低的资源池.通过支持向量机选择空闲的节点,从而达到云节点选择的多维性,同时兼顾了云计算的自适应问题,使得分布在云上的资源能满足日益增长的性能需要.

1 算法及模型的理论依据及设计思想

1.1 支持向量机

支持向量机理论是统计学习范畴的一个重要分支,特别适合解决云计算、非线性和高维模式识别问题中表现出特有的优势.对于训练集 $\{x_i, y_i\}$, (其中 x 是输入向量, y 是类别标签), 存在一个超平面,使得两类样本数据完全分开, $w^T x + b = 0$ ^[2].

1.2 归类问题

假设待支持向量机的云资源集合用 N 来表示,置信的效率云计算集合用 D 表示,支持向量机的结果集合用 B 表示,从而得出 B 就是一由个 0~1 数字元素组成的布尔值变量的集合,用 1 表示完全符合的云资源,0 表示完全拒绝的云资源,云效率模型识别的结果依据于识别器 D 到检测结果 B 的 Vague 关系集合,这将同时反映模糊检验的肯定以及否定的程度,和它们之间的不确定的因素.

整个分类问题可以化作计算如下公式:

$$\gamma = \frac{1}{2} \left[\left(\frac{\omega}{\|\omega\|} \cdot \chi^+ \right) - \left(\frac{\omega}{\|\omega\|} \cdot \chi^- \right) \right]$$

$$\gamma = \frac{1}{2 \|\omega\|} (\langle \omega, x^+ \rangle - \langle \omega, x^- \rangle) = \frac{1}{\|\omega\|} \quad [3]$$

如果该等式成立,即表示线性可分的情况下,选择最优平面作为线性规划问题的拉格朗日解.式中: $d_j \in D$, $b_k \in B$.模糊检验的基础是首先选择理论上信赖的云来进行扩展,逐步得到一个效率云集合,这个效率的云集合可以作为检验的知识

库,最终用核函数关系表示待检测云与这个效率云集合的关系,集合的元素值就是待检测云与知识库的可信度因素,从而实现了变化极快并且极复杂的云资源信息效率因素的表达.在上述的建模过程中,采用的是常规的 max 和 min 来合成运算,也可以用其他集合运算来表达,例如计算贴近度,计算统计汉明距离等 $d(x) = (w^*)^T x + b^*$ 来确定 x 的归类.

1.3 云计算调度机制

MapReduce 是 Google 提出的一个软件框架它对大数据的处理方式采用分布式的模式,即: Map 和 Reduce^[4].将大数据进行切分块,将非数据集分割成小数据集针对不同的云节点进行处理,从而实现并行算法,每个元数据就是一系列的<键,值>对,数据的预处理过程抽象成 Map 映射和 Reduce 两大阶段.在 MapReduce 平台里,需要将复杂的多维云节点资源划分为线性单元,由 Map 处理器处理,将一组<键,值>对映射成一组新的<键,值>对,输入到 Reduce 单元.MapReduce 的计算过程如图 1 所示.

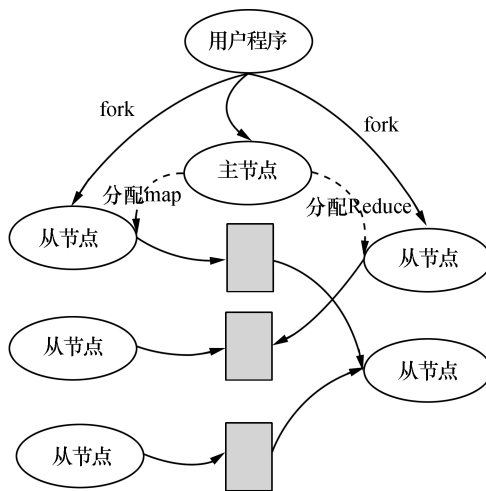


图 1 MapReduce 算法模型

Fig.1 MapReduce algorithm model

2 算法的具体实现及新的任务调度模型

当前的 Mapreduce 算法虽然考虑了并行的思想,但是智能性和效率不高,需要大量处理器才能精准的进行分配调度^[5].针对这个突出问题,提出了一种基于支持向量机的 MapReduce 的任务调度算法.

2.1 运用 SVM 方法来对调度节点线性可分

通过前面支持向量机的基础定义,我们找到了将支持向量机应用到调度任务的一个途径,即通过把待分配的子节点 SLOT 分为一个集合,形成一个模糊的识别器,并通过划分以及随机组合,形成每个元素为 3 的子集合 $\{D_1, D_2, D_m\}$, 其中 $n=c_m^3$, n 是每个识别器的数量.检测结果集合 = $\{1, 0\}$, 假定云计算的检验知识集就是一个定义在集合与集合之间线性可分关系, 设为: $R = \{R_1, R_2, \dots, R_n\}$. 调度模型如图 2 所示.

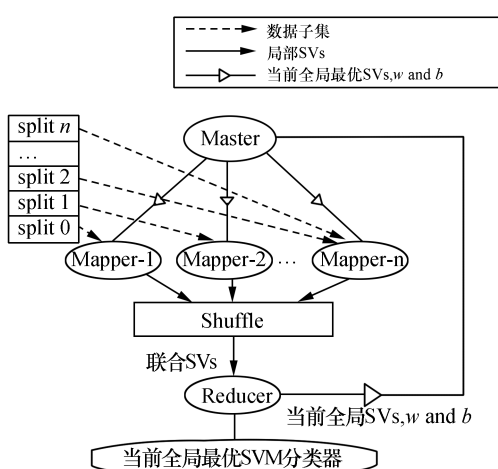


图 2 基于 SVM 的任务调度模型

Fig.2 Task scheduling model based on SVM

基于支持向量机的任务调度过程分为以下步骤:

输入:待分配非线性云节点维度 D , 节点个数 P , 阈值设置为 g .

1) Split $D \rightarrow P$, 定义为 D_i 将 D_i 分发到每个 Slot 主节点上.

2) 参数初始化操作, 为每个样本初始化参数 g .

3) 循环执行 Dgap, 直到大于阈值退出.

4) 选择待分配节点和 SLOT 的 Dgap 值.

5) 根据 4) 的返回结果值, 得出最终的拉格朗日的云计算值. 将整个任务分成 P 个子任务, Mapper 函数中通过核函数找出每个子任务的上限和下限值, 然后将所有子任务上限任务选出特定样本节点, 分别对应所有最小值, 选出所有子任务中对应 g 的最大的, 然后使用核函数更新两个样本的 α .

6) 将步骤五中作为该任务的输入, 在 Mapper 函数中更新两个样本对应的 f .

2.2 基于支持向量机的 MapReduce 任务调度模型

模型通过对云节点的特征向量对任务调度模型进行了建模, Hadoop^[6-11] 平台下的 MapReduce 任务调度主要针对的是线性节点, 基于 Map 以及 Reduce 并行化处理.

输出 待分配节点

Procedure

Begin

R = number of Cloud

收到计算节点任务 Cn 请求

I = 1;

Do Begin:

建立用户的云请求到集合 D_i 的匹配度核函数分析;

从待检测的云计算库中找出与 D_i 对应的检测依据核函数分析;

对该集合进行 max 和 min 合成运算得到检测结果的核函数分析 B ;

通过计算得到核函数 S 和精确检测函数 H

如果 检测结果 = 1 那么

Map(key, value); Reduce(key, value);

否则

I = I + 1;

End while(i > n) /* 取决云服务提供商提供的指定云资源数量 */

结束

3 实验分析

3.1 数据集

在相邻两周数据的矛盾率为 0.14 的条件下, 本文基于支持向量机实现了由调度属性仿真模拟机 GDW325, 在实验环境下分别向调度器总控提交不同类型的作业申请, 随时在调度任务执行过程中对效能特征值进行收集统计, 统计区设置为 1 s, 每 1 s 收集约 46 个云计算节点, 并根据这些节点的特征属性, 建立一套适应 SVM 的样本数据集.

3.2 结果分析

支持特征属性云计算性能的归约针对数字型的调度节点特征值采用非线性方法化操作. 设置属性阈值为 g , 得到如图 3 所示的任务调度耗时大于 0.142 的性能特征曲线, 如图 4 所示集群系统中负载的能耗与调度任务节点的各个组成部件均有相关性, 其中随着特征向量的维度的增加, 所表现出的执行支持向量机执行重要程度就越大.

我们通过仿真实验,得出三组不同的实验结果.这是由于任务调度分配过程中系统产生的调度消耗主要来自于处理寻找相应的云节点.新的模型针对非线性维特征对分配调度的性能影响提出了改善,特别是对于多维以及非线性云节点较之传统的 Kmeans 算法有明显的提高.SVM 最擅长的就是对分线性云节点的处理,因此算法对于作业的选择分配速度明显高于 Kmeans 算法.而其他单分配任务算法仅考虑了线性云节点对负载能耗的影响,在多维云计算环境下,SVM 处理多维的能力就表现得更突出.

根据分析,通过基于 SVM 算法的 MapReduce 并行处理的优势更加明显.

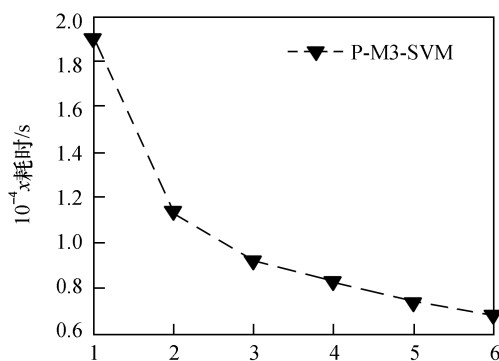


图 3 单作业任务调度模型耗时分析

Fig.3 Time consuming analysis of single job task scheduling model

从图 3 可以看出,随着云节点规模的不断扩大,数据处理的时间明显增多,可以得出,对于多维非线性云节点的分配问题,随着节点种类的增多,基于 SVM 的任务调度模型的处理能力不会下降,从而有效的验证了基于支持向量机的任务调度模型有良好的扩展性.

由图 4 和图 5 可以看出,User3 是相对正常的云资源,并且 User3 可以适用于更多云资源的迭代运算,也就是说,随着云服务资源的增加,通过上述算法模型,可以计算出更加准确的云资源来满足用户的效率需求.为了分析时间性能和效率,仿真实验比较了线性算法和支持向量机算法,由实验结果可知,SVM 在 MapReduce 任务调度上可以有效的减少任务调度选择节点的时间,提高了 Map&Reduce 的速度,适合处理大规模的,多样存储的数据样本,虽然在对处理器要求上低于一般 Hadoop,但是仍然保持了较高的分类精度.

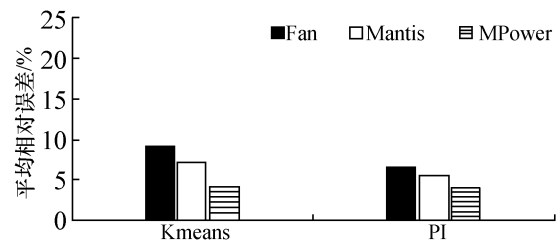


图 4 单作业任务调度模型估计与 Kmeans 算法比较

Fig.4 Comparison of single job task scheduling model and Kmeans algorithm

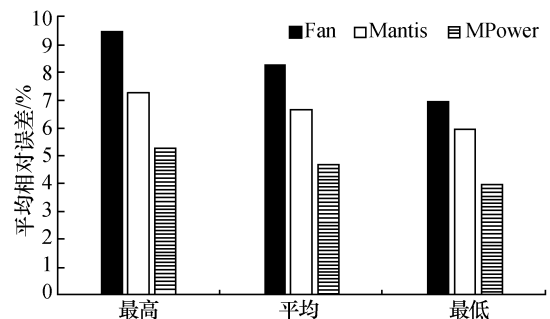


图 5 基于分类精度多作业任务调度模型比较图

Fig.5 Comparison chart based on the classification accuracy of multi task scheduling model

4 结束语

本文引入支持向量机的相关机制到 MapReduce 任务调度的设计中,提出了一种支持向量机的任务调度模型.新模型对待分配云节点进行识别,满足了云资源分配的多维性,高效性.加强了云计算的效率环境,目前该模型的研究工作仍在进行之中,尚有许多需要改进的地方,如:云计算系统的自适应性,智能性等.

参考文献:

- [1] ZADEH L A. Responses to Elkan why the success of fuzzy logic is not paradoxical[J]. IEEE Expert, 2003, 9 (4): 43-46.
- [2] DEAN J, GHEMAWAT S. MapReduce: Simplified data processing on large clusters[J]. Proceedings of the Conference on Operating System Design and Implementation (OSDI), 2008, 51(2): 107-113.
- [3] BURROWS M. The Chubby lock service for loosely-coupled distributed systems[J]. Proceedings of the 7th symposium on Operating systems design and implementation, 2006, 6(2): 335-350.

(下转第 95 页)

- [4] 卢胜进,邓国增,李爱华.江永香姜产业现状及发展对策研究[J].湖南农业科学,2009(12):98-101.
- [5] 王连珠,周昱,陈泳,等.QuEChERS 样品前处理-液相色谱-串联质谱法测定蔬菜中 66 种有机磷农药残留量方法评估[J].色谱,2012,30(2):146-153.
- [6] 叶雪珠,赵燕申,王强,等.蔬菜农药残留现状及其潜在风险分析[J].中国蔬菜,2012(14):76-80.
- [7] 杜娟.果蔬中有机磷农药残留检测方法的研究进展[J].广东化工,2012,39(9):149-150.
- [8] 田憬若,陈文,张长江,等.不同前处理方法对蔬菜中有机磷农药残留检测的影响[J].现代食品科技,2013,29(3):664-667.
- [9] 惠瑾,王玮,刘超,等.浅谈农药残留检测的前处理技术研究进展[J].农业装备技术,2013,39(5):7-9.
- [10] 孙春燕,李宏坤,平红,等.AuNPS/Sol-gel 复合膜法固定乙酰胆碱酯酶生物传感器检测有机磷农药[J].高等学校化学学报,2011,32(11):2533-2538.
- [11] 杨东顺,梅文泉,汪禄祥,等.蔬菜中 16 种有机磷农药残留量的气相色谱快速测定[J].西南农业学报,2015,28(5):2136-2141.
- [12] LI Y P, HAN G Y. Ionic liquid-functionalized graphene for fabricating an amperometric acetylcholinesterase biosensor[J]. Analyst, 2012, 137(13):3160-3165.
- [13] TUZIMSKI T, SOCZEWINSKI E. Correlation of retention parameters of pesticides in normal and reversed-phase systems and their utilization for the separation of a mixture of 14 triazines and urea herbicides by means of two-dimensional thin layer chromatography[J]. Journal of chromatography A, 2012, 961(2):277-283.
- [14] LI Y, DONG F, LIU X, et al. Simultaneous enantioselective determination of fenbuconazole and its main metabolites in soil and water by chiral liquid chromatography / tandem mass spectrometry[J]. J Chromatogr A, 2011, 1218(38):6667-6674.
- [15] CAI L S, XING J, DONG L, et al. Ionic liquid-functionalized graphene for fabricating an amperometric acetylcholinesterase biosensor[J]. Analyst, 2012, 137(13):3160-3165.
- [16] HU Y Y, ZHENG P, HE Y Z, et al. Direct electrochemistry of catalase at amine-functionalized graphene/gold nanoparticles composite film for hydrogen peroxide sensor[J]. Electrochimica acta, 2011, 56(7):2957-2953.

(上接第 84 页)

- [4] BERLINSKA J, DROZDOWSKI M. Scheduling divisible MapReduce computations[J]. Journal of Parallel and Distributed Computing, 2011, 71(3):450-459.
- [5] SANDHOLM T, LAI K. Dynamic proportional share scheduling in hadoop Job scheduling strategies for parallel processing[J]. Dynamic proportional share scheduling in hadoop Job scheduling strategies for parallel processing, 2010, 62(2):110-131.
- [6] KAMBATLA K, PATHAK H, PUCHA H. Towards optimizing hadoop provisioning in the cloud[J]. Proc of the First Workshop on Hot Topics in Cloud Computing, 2009, 3(2):118-118.
- [7] ZAHARIA M. Job scheduling for multi-user mapreduce[J]. Job scheduling for multi-user mapreduce clusters, 2009, 55(3):117-121.
- [8] DOELITZSCHER F. Private cloud for collaboration and e-Learning services; from IaaS to SaaS Computing[J]. Private cloud for collaboration and e-Learning services; from IaaS to SaaS Computing, 2011, 91(2):23-42.
- [9] WANG C, WANG Q, Ren K. Privacy-preserving public auditing for data storage security in cloud computing[J]. In Proceedings of IEEE INFOCOM, 2010, 3(2):59-60.
- [10] KIM J, BENTLEY P J. Towards an Artificial Immune System for Network Intrusion Detection: An investigation of Dynamic Clonal Selection with negative Selection Operator[J]. Proceedings of the Congress on Evolutionary Computation(CEC-2002), 2002, 17(5):1015-1020.
- [11] MAJI P K, BISWAS, ROY A R. Fuzzy soft sets[J]. Journal of Fuzz Mathematics, 2001, 9(3):58-60.