(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105550374 A
(43)申请公布日 2016.05.04

(21)申请号 201610069166.8

(22)申请日 2016.01.29

(71)申请人 湖南大学
地址 410082 湖南省长沙市岳麓区麓山南路1号

(72)发明人 唐卓 陈建国 李肯立 鲁彬 陈俊杰 肖锦波

(74)专利代理机构 深圳市兴科达知识产权代理有限公司 44260
代理人 王翀

(51)Int.Cl.
G06F 17/30(2006.01)
G06N 99/00(2010.01)

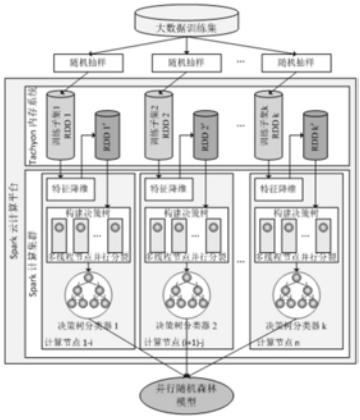
权利要求书2页 说明书5页 附图2页

(54)发明名称

Spark 云服务环境下面向大数据的随机森林并行机器学习方法

(57)摘要

本发明公开了一种 Spark 云服务环境下面向大数据的随机森林并行机器学习方法，通过特征向量重要性分析对高维大数据降维处理，并采用加权投票方式进行预测；利用分布式内存管理机制、云计算平台，对随机森林训练过程模型构建、单棵决策树分裂过程以及预测投票等三层并行化进行改进。本发明通过特征向量重要性分析对高维大数据降维处理和采用加权投票方式进行预测，实现了随机森林的方法优化，提高了随机森林机器学习方法对复杂大数据的挖掘效果；在此基础上进行基于 Spark 云平台的随机森林并行化方法，提高了随机森林机器学习方法的运行效率。



CN 105550374 A

1.一种Spark云服务环境下面向大数据的随机森林并行机器学习方法,其特征在于,包括如下步骤:

步骤一、使用随机森林模型进行训练过程得到训练完成的随机森林模型;训练过程中使用特征重要性分析方法进行训练集数据的特征降维处理;使训练集数据的特征从M维降低为m维,m<M;

步骤二、使用训练完成的随机森林模型对测试数据集进行预测过程得到预测结果;预测过程中使用特征重要性分析方法进行测试数据集数据的特征降维处理;使测试数据集数据的特征从M维降低为m维;

步骤三、将训练完成的随机森林模型中各个决策树模型的训练准确度作为预测投票的权值,对预测结果进行加权投票,得到最终的预测结果。

2.如权利要求1所述的Spark云服务环境下面向大数据的随机森林并行机器学习方法,其特征在于,所述步骤一包括如下步骤:

1.1、对数据的训练集进行放回抽样生成n个训练数据子集;n为自然数;使用随机森林模型对训练数据子集进行训练,在不同计算节点之间创建n个映射(Map)任务完成对训练数据子集的训练;

1.2、计算每棵决策树分裂过程中每一个特征变量的信息增益;

1.3、计算训练数据子集中每个特征变量的信息熵;

1.4、计算训练数据子集中每个特征变量的自信息;

1.5、计算出每个特征变量的信息增益率;

1.6、对训练数据子集的所有特征变量的重要性值进行降序排列,所述特征变量的重要性值为训练数据子集的特征变量的信息增益率 占训练数据子集的全部特征变量的信息增益率之和的比例;选取前k个特征变量;再从剩下的M-k个特征中随机挑选m-k个特征,共同组成m个特征,将数据从M维降低为m维;得到训练完成的随机森林模型;

其中,M为数据降维前的维数;m为数据降维后的维数;K为自然数,k＜m,m＜M。

3.如权利要求2所述的Spark云服务环境下面向大数据的随机森林并行机器学习方法,其特征在于,所述步骤二包括如下步骤:

2.1、计算训练完成的随机森林模型中各个决策树的各个特征变量的信息增益率;

2.2、计算训练完成的随机森林模型中全部训练数据子集的各个特征变量的信息增益率;

2.3、计算测试数据集各个特征变量的测试数据集重要性值,测试数据集重要性值指:在测试数据集中每个测试数据集的特征变量的信息增益率占测试数据集的全部特征变量的信息增益率的比例;

2.4、对测试数据集的各个特征变量的测试数据集重要性值进行降序排列,选取测试数据集的前m个重要性值,将高维的测试数据集从M维降低为m维,其中,m＜M。

4.如权利要求1所述的Spark云服务环境下面向大数据的随机森林并行机器学习方法,其特征在于,使用Tachyon分布式内存处理平台将数据加载到云服务平台中,利用分布式内存管理机制、云计算平台,使用机器学习方法进行随机森林模型在训练过程中的构建、单棵决策树分裂过程和预测投票过程,并分别进行并行化处理。

5.如权利要求4所述的Spark云服务环境下面向大数据的随机森林并行机器学习方法,

其特征在于,所述云服务平台为Spark云服务平台;数据以弹性分布式数据集对象的形式存储在Spark平台的Tachyon分布式内存系统中。

Spark云服务环境下面向大数据的随机森林并行机器学习方法

技术领域

[0001] 本发明属于计算机领域，尤其涉及一种Spark云服务环境下面向大数据的随机森林并行机器学习方法。

背景技术

[0002] 术语解释：

[0003] 特征降维：进行图像或数据的特征提取的过程中，提取的特征维数太多经常会导致特征匹配时过于复杂，消耗系统资源，此时采用一个低纬度的特征来表示高纬度即为特征降维。

[0004] 随着各种新型信息发布方式的不断涌现，以及云计算、物联网等技术的兴起，以及遍布地球各个角落的各种各样的传感器，数据正以前所未有的速度在不断地增长和累积，大数据时代已经来到。随着网络应用日益深化，大数据应用的价值越来越明显。海量数据里面蕴含着大量十分有价值的数据，要处理的数据量越来越大、而且还将更加快速地增长，同时业务需求和竞争压力对数据处理的实时性、有效性也提出了更高要求。传统的常规数据处理技术已无法应付，大数据带来了很多现实的难题。如何通过机器学习、数据挖掘等方法从这些大数据中更快速、更精确地挖掘出有价值的数据，是当今学术界和工业界研究的热点。为了解决这些难题，我们需要突破传统技术，根据大数据的特点进行新的技术变革。

[0005] 在基于云计算平台的分布式数据挖掘方向的研究，已经得到了广泛地展开并且取得了大量优秀的成果。Hadoop是目前大数据平台中 应用率最高的技术，特别是针对诸如文本、社交媒体订阅以及视频等非结构化数据。MapReduce采用shared-nothing架构设计，在执行Job时，各个Job之间是彼此隔离的，只能通过HDFS等部件进行交互，因此在各个任务间的全局同步或者状态的共享是一个很大的挑战。而在MapReduce在处理过程中，会将map中间的结果写入本地磁盘，然后再通过shuffle机制发送至reduce进行处理，因此也不适合需要大量的网络通讯任务。另外，MapReduce是一种批量处理架构，这也意味着它并不适合于实时或流式的数据访问，在处理联机事务处理(OLTP,OnLine Transaction Processing)型任务时也显得力不从心。如何能避免MapReduce所带来的问题，同时又能充分利用其优越的海量数据处理能力，成为实践中重要的问题。

[0006] Spark是由UC Berkeley AMPLab实验室于2009开发的开源数据分析集群计算框架，是Berkeley Data Analytics Stack(BDAS)中的核心项目，被设计用来完成交互式的数据分析任务。Spark提供了比Hadoop更为通用和灵活的操作接口。与Hadoop提供的Map和Reduce相比，Spark基于RDD抽象，提供的数据集操作类型更多。Spark允许将一份数据缓存在内存中，在同一份数据上迭代计算，因此，Spark更适合于迭代运算较多的机器学习或数据挖掘运算。RDD可以把数据cache到内存中，下一步操作直接从内存中输入，省去了MapReduce大量的磁盘IO操作，这对于迭代运算比较常见的机器学习方法来说，效率提升会相当大。

[0007] 随机森林是一种集成学习的方法，它在高维大数据上面表现出不 俗的效果。随机森林机器学习方法采用了特征子空间来构建模型，当数据中的噪声过多时，随机森林构建的分类器可能会包含噪声，集成这些噪声分类器进行分类预测可能会降低随机森林机器学习方法的整体分类效果。而将随机森林机器学习方法并行化可以提高随机森林机器学习方法的执行速度。

[0008] 传统的分类方法在低维的小数据集上面可以取得比较理想的效果，但是当数据的结构变得复杂，数据的维度变高，数据的大小增大时，传统的分类方法的性能则会明显地下降。面对海量的大数据，传统分类方法在建模和预测的过程需要花费比较多的时间。因此，如何选择适合的模型，使随机森林机器学习方法在低维和高维的数据集上都能拥有较好的分类性能成为本发明的重点研究问题。

发明内容

[0009] 为解决上述问题，本发明提供了一种Spark云服务环境下面向大数据的随机森林并行机器学习方法。本发明通过特征向量重要性分析对高维大数据降维处理，并采用加权投票方式进行预测，从以上两个方面实现随机森林的方法优化，有效提高随机森林机器学习方法对复杂大数据的挖掘效果；为了提高该方法的性能，在此基础上提出了基于Spark云平台的随机森林并行化方法，利用分布式内存管理机制、云计算平台，通过对随机森林训练过程模型构建、单棵决策树分裂过程以及预测投票等三层并行化进行改进，提高了随机森林机器学习方法的运行效率。

[0010] 为达到上述技术效果，本发明的技术方案是：

[0011] 一种Spark云服务环境下面向大数据的随机森林并行机器学习方 法，包括如下步骤：

[0012] 步骤一、使用随机森林模型进行训练过程得到训练完成的随机森林模型；训练过程中使用特征重要性分析方法进行训练集数据的特征降维处理；使训练集数据的特征从M维降低为m维，m<M；

[0013] 步骤二、使用训练完成的随机森林模型对测试数据集进行预测过程得到预测结果；预测过程中使用特征重要性分析方法进行测试数据集数据的特征降维处理；使测试数据集数据的特征从M维降低为m维；

[0014] 步骤三、将训练完成的随机森林模型中各个决策树模型的训练准确度作为预测投票的权值，对预测结果进行加权投票，得到最终的预测结果。

[0015] 进一步的改进，所述步骤一包括如下步骤：

[0016] 1.1、对数据的训练集进行放回抽样生成n个训练数据子集；n为自然数；使用随机森林模型对训练数据子集进行训练，在不同计算节点之间创建n个映射(Map)任务完成对训练数据子集的训练；

[0017] 1.2、计算每棵决策树分裂过程中每一个特征变量的信息增益；

[0018] 1.3、计算训练数据子集中每个特征变量的信息熵；

[0019] 1.4、计算训练数据子集中每个特征变量的自信息；

[0020] 1.5、计算出每个特征变量的信息增益率；

[0021] 1.6、对训练数据子集的所有特征变量的重要性值进行降序排列，所述特征变量的

重要性值为训练数据子集的特征变量的信息增益率占训练数据子集的全部特征变量的信息增益率之和的比例;选取前k个特征变量;再从剩下的M-k个特征中随机挑选m-k个特征,共同组 成m个特征,将数据从M维降低为m维;得到训练完成的随机森林模型;

[0022] 其中,M为数据降维前的维数;m为数据降维后的维数;K为自然数,k＜m,m＜M。

[0023] 进一步的改进,所述步骤二包括如下步骤:

[0024] 2.1、计算训练完成的随机森林模型中各个决策树的各个特征变量的信息增益率;

[0025] 2.2、计算训练完成的随机森林模型中全部训练数据子集的各个特征变量的信息增益率;

[0026] 2.3、计算测试数据集各个特征变量的测试数据集重要性值,测试数据集重要性值指:在测试数据集中每个测试数据集的特征变量的信息增益率占测试数据集的全部特征变量的信息增益率的比例;

[0027] 2.4、对测试数据集的各个特征变量的测试数据集重要性值进行降序排列,选取测试数据集的前m个重要性值,将高维的测试数据集从M维降低为m维,其中,m＜M。

[0028] 进一步的改进,使用Tachyon分布式内存处理平台将数据加载到云服务平台中,利用分布式内存管理机制、云计算平台,使用机器学习方法进行随机森林模型在训练过程中的构建、单棵决策树分裂过程和预测投票过程,并分别进行并行化处理。

[0029] 进一步的改进,所述云服务平台为Spark云服务平台;数据以弹性分布式数据集(RDD)对象的形式存储在Spark平台的Tachyon分布式内存系统中。

[0030] 本发明的优点如下:

[0031] 1.本发明针对大数据具有高维特征的问题,分别在训练过程和预测过程使用特征重要性分析的方法进行高维数据的特征降维处理,有效降低了方法的计算量和复杂度;针对大数据中存在大量噪声数据问题,采用加权投票的方式进行数据集预测和投票,降低含噪声数据比率较高的决策树分类器投票权重,提高含噪声数据比率较低的决策树分类器投标权重,减少噪声数据对数据分类投票结果的影响,提高随机森林机器学习方法对复杂大数据的分类准确度。

[0032] 2.本发明在提高随机森林机器学习方法对复杂大数据的分类准确度的同时,将面向大数据的随机森林改进方法在Spark云平台中进行并行化实现,利用分布式内存管理机制、云计算平台,对随机森林训练过程模型构建、决策树训练过程,预测投票等三层并行化进行改进,提高随机森林机器学习方法的运行效率。

附图说明

[0033] 图1为本发明所述方法的特征选择和降维过程示意图;

[0034] 图2为本发明所述方法新型随机森林的结构图。

[0035] 图3为本发明所述方法的基于Spark云服务环境下的新型随机森林并行机器学习方法结构图。

具体实施方式

[0036] 下面将结合附图和实施例对本发明做进一步的说明。

[0037] 实施例1

[0038] (1)针对大数据具有高维特征的问题,分别在训练过程和预测 过程使用特征重要性分析的方法进行高维数据的特征降维处理,在效降低方法的计算量和复杂度。针对大数据中存在大量噪声数据问题,采用加权投票方式进行数据集预测和投票,减少噪声数据对数据分类投票结果的影响,提高随机森林机器学习方法对复杂大数据的分类准确度。

[0039] 步骤1:随机森林模型训练过程中对训练数据的特征选择过程,其过程如图1所示。具体实现步骤如下:

[0040] 步骤1.1:对高维大数据训练集进行有放回的抽样成n个训练数据子集;

[0041] 步骤1.2:计算每棵决策树分裂过程中每一个特征变量的信息增益;

[0042] 步骤1.3:计算该样本子集中每个特征变量的信息熵;

[0043] 步骤1.4:计算训练样本集中每个特征变量的自信息;

[0044] 步骤1.5:计算每个特征变量的特征变量的信息增益,为了克服训练过程中产生过似合现象,即使用信息增益选择特征变量时偏向选择取值较多的特征变量的问题,在此使用信息增益率来选择特征变量;

[0045] 步骤1.6:最后,对各个特征变量的重要性值进行降序排列,并选取前k(k<<M,k<m)个重要性值最大的特征变量,然后从剩下的M-k个特征中随机挑选(m-k)个特征。共同组成m个特征,将高维数据从M维降低为m维。这里的特征重要性是指:在一个训练子集中,每个特征变量的重要性是指该特征变量的信息增益率占全部特征变量的信息增益率的比例。

[0046] 步骤2:在数据预测过程中的面向高维大数据的特征降维过程,具体实现步骤如下:

[0047] 步骤2.1:在对训练样本数据进行随机森林中的各个决策树训练过程完成之后,计算各个决策树的各个特征变量的加权信息增益率。

[0048] 步骤2.2:计算整个随机森林模型中,全部训练样本数据的各个特征变量的加权信息增益率。

[0049] 步骤2.3:计算各个特征的特征重要性,在训练集中,每个特征变量的重要性是指该特征变量的信息增益率占全部特征变量的信息增益率的比例。

[0050] 步骤2.4:对各个特征变量的重要性值进行降序排列,选取测试数据集的前m(m<<M)个重要性值最大的特征变量,将高维的测试数据集从M维降低为m维。

[0051] 步骤3:使用经过训练的随机森林模型对待测试数据进行预测,然后将随机森林中各个决策树模型的训练准确度作为其预测投票的权值,对预测结果进行加权投票,得到最终的预测结果。图2为新型随机森林并行机器学习方法的结构设计。

[0052] (2)为了提高本专利所提出的新型随机森林机器学习方法的运算性能,在此使用Spark云服务平台对该方法进行并行化实现。Spark云服务环境采用10台计算机节点组成,包括1台主节点和9台从节点。每台计算机节点的配置为Intel Quad Core 2.66GHZ CPU,8GB内存、Centos 5.6Linux操作系统。所有计算机都通过高速光纤网络互连。Apache Spark软件版本为1.1.0,方法采用R语言实现。

[0053] 图3为本专利所公开的Spark云服务环境下面向大数据的新型随机森林并行机器学习方法原理图。具体实现步骤说明如下:

[0054] 步骤1:加载大数据到Spark平台,在对大数据进行训练、预测和投票之前,需要先将这些数据加载到Apache Spark平台中。我们将这些大数据集以RDD对象的形式存储在

Spark平台的Tachyon内存系统中。

[0055] 步骤2:随机森林训练过程中的并行化模型构建,当训练数据集加载到Tachyon系统之后,训练数据集被抽样成k个训练子集。在随机森林模型训练过程中,我们将在不同计算节点之间创建k个Map任务,用于完成这k个训练子集的模型训练任务。这k个Map任务将并行执行。

[0056] 步骤2.1:在Map阶段,k个训练子集所对应的k棵决策树分类器将被训练构建。各棵决策树的分类准确度CAi(x)也会在partition阶段通过对各个训练子集的袋外数据集OOB测试结果计算得到。这些中间结果都被以RDD对象的方式存储在Tachyon内存系统中。

[0057] 步骤2.2:在Reduce阶段,各个决策树分类器模型hi(x)及其分类准确度CAi(x)将被合并计算,并返回最终的随机森林模型。

[0058] 步骤3:加载测试数据到Spark平台的Tachyon内存系统中。

[0059] 步骤4:在预测和投票之前,首先需要将已经训练完成的随机森林模型部署到Spark平台的相应计算节点上。

[0060] 步骤5:针对每个测试数据,每一个测试数据都需要经过随机森林模型的k个决策树分类器进行预测,并产生相应的预测结果。本步 骤对这k个预测过程进行并行化,使k棵决策树同时在k个计算节点中进行预测。

[0061] 以上实例的说明只是用于帮助理解本发明的核心思想;同时,对于本领域的一般技术人员,依据本发明的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本发明的限制。

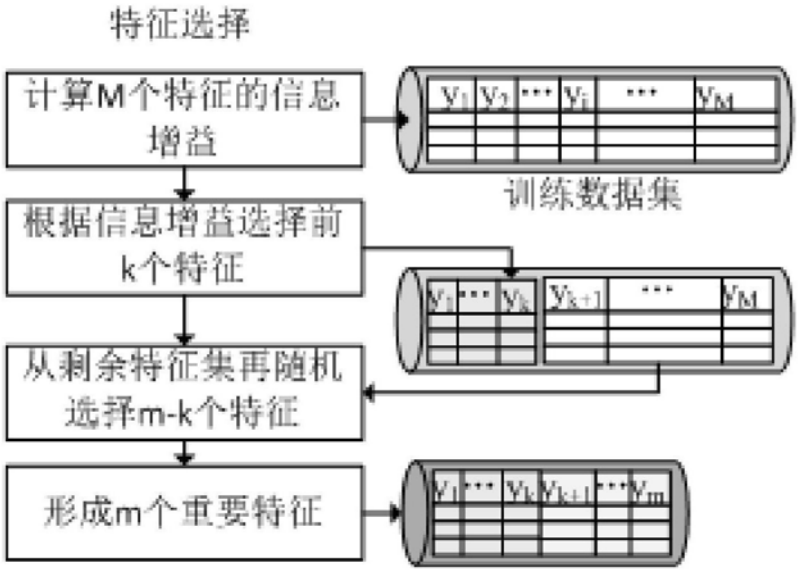

特征选择
计算M个特征的信息增益
训练数据集
根据信息增益选择前k个特征
从剩余特征集再随机选择m-k个特征
形成m个重要特征

图1

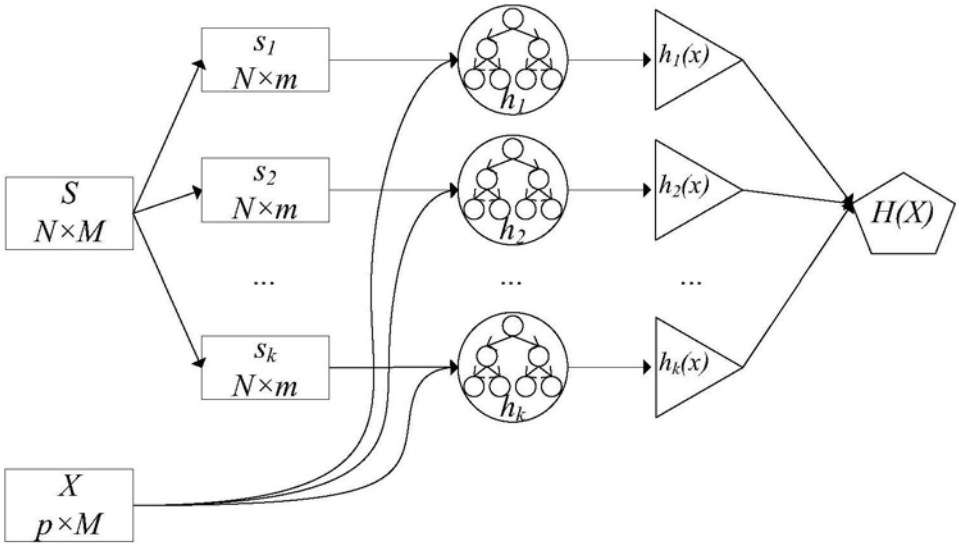

S
N×M
X
p×M
H(X)

图2

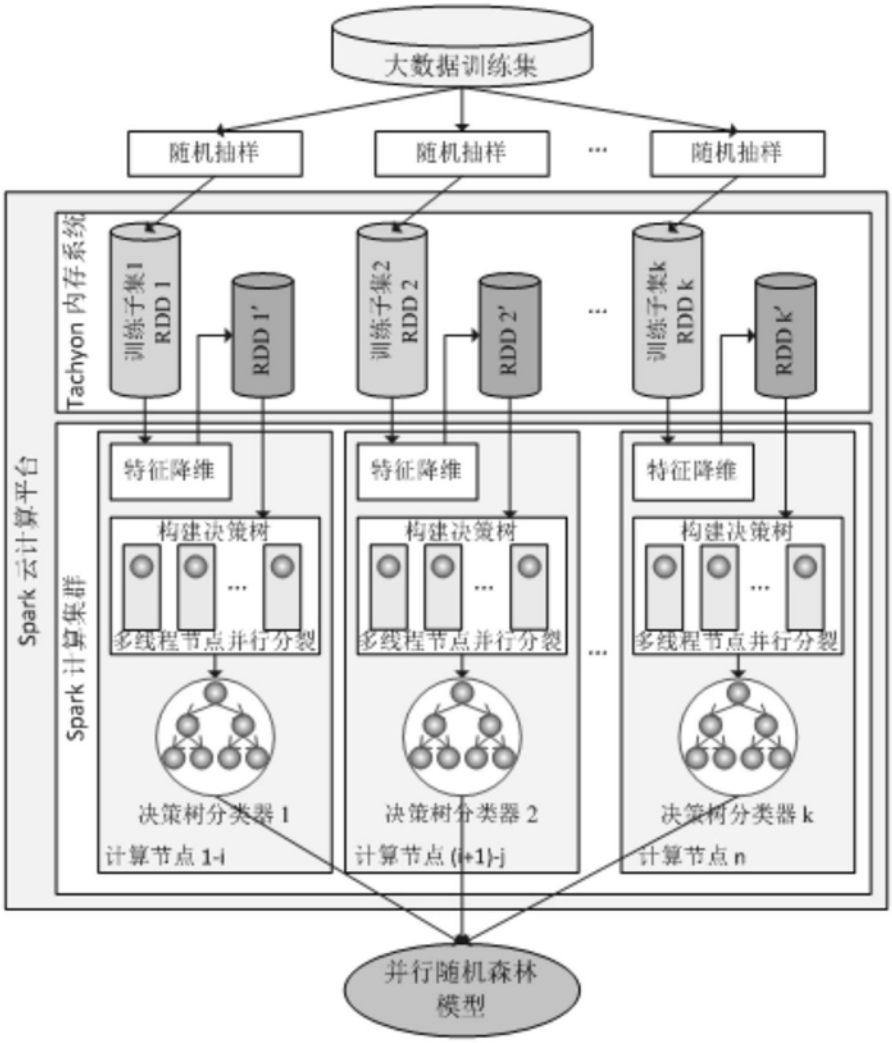

大数据训练集
随机抽样
随机抽样
...
随机抽样
Tachyon 内存系统
训练子集1 RDD 1
RDD 1'
训练子集2 RDD 2
RDD 2'
...
训练子集k RDD k
RDD k'
Spark 云计算平台
Spark 计算集群
特征降维
特征降维
特征降维
构建决策树
构建决策树
构建决策树
多线程节点并行分裂
多线程节点并行分裂
多线程节点并行分裂
决策树分类器 1
决策树分类器 2
决策树分类器 k
计算节点 1-i
计算节点 (i+1)-j
计算节点 n
并行随机森林模型

图3