(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2017 年 10 月 5 日 (05.10.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/167097 A1

(51) 国际专利分类号:
G06F 21/62 (2013.01)

(21) 国际申请号: PCT/CN2017/077704

(22) 国际申请日: 2017 年 3 月 22 日 (22.03.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610201626.8 2016 年 3 月 31 日 (31.03.2016) CN

(71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛大开曼资本大厦一座四层 847 号邮箱, Grand Cayman (KY)。

(72) 发明人; 及

(71) 申请人 (仅对美国): 姜晓燕 (JIANG, Xiaoyan) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 王少萌 (WANG, Shaomeng) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。 杨旭 (YANG, Xu) [CN/CN]; 中国浙江省杭州市余杭区文一西路 969 号 3 号楼 5 楼阿里巴巴集团法务部, Zhejiang 311121 (CN)。

(74) 代理人: 北京国昊天诚知识产权代理有限公司 (CO-HORIZON INTELLECTUAL PROPERTY INC.); 中国北京市朝阳区小关北里甲 2 号渔阳置业大厦 B 座 605, Beijing 100029 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH,

[见续页]

(54) Title: METHOD AND APPARATUS FOR TRAINING MODEL BASED ON RANDOM FOREST

(54) 发明名称 : 一种基于随机森林的模型训练方法和装置

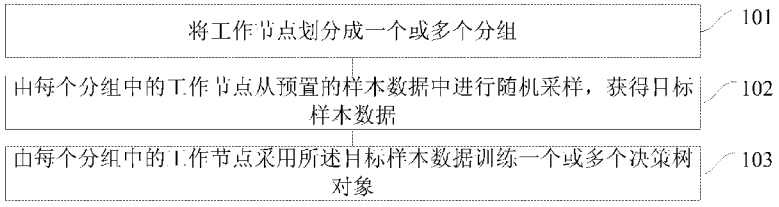


图 1

101 Divide worker nodes into one or more groups
102 Worker nodes from each group randomly sample preset sample data, so as to obtain target sample data
103 Worker nodes from each group use the target sample data to train one or more decision tree objects

(57) Abstract: Disclosed are a method and an apparatus for training a model based on a random forest. The method comprises: dividing worker nodes into one or more groups (101); worker nodes of each group randomly sampling preset sample data so as to obtain target sample data (102); worker nodes of each group using the target sample data to train one or more decision tree objects (103). The method of the present invention has the advantage that it is not necessary to completely scan all sample data at one time, reducing greatly the volume of data to be read, thereby reducing time needed for iterative updating of a model and improving training efficiency.

(57) 摘要: 一种基于随机森林的模型训练方法和装置，该方法包括：将工作节点划分成一个或多个分组（101）；由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据（102）；由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象（103）。该方法不需要扫描一次完整的样本数据，大大降低了数据的读取量，减少了耗费的时间，进而减少模型的迭代更新时间、提高训练效率。

WO 2017/167097 A1

CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

一种基于随机森林的模型训练方法和装置

技术领域

本申请涉及计算机处理的技术领域，特别是涉及一种基于随机森林的模型训练方法和一种基于随机森林的模型训练装置。

背景技术

随着互联网的快速发展，人们生活的方方面面都与互联网产生了联系，在人们使用互联网的相关功能时，产生了海量的数据。

目前，经常使用随机森林（Random forest）算法进行模型训练，对这些海量的数据进行挖掘，从而进行分类、推荐等操作。

随机森林是一个树型分类器{h(x,k),k=1,...}的集合，元分类器 h(x,k)一般是用 CART（Classification And Regression Tree，分类回归树）算法构建的没有剪枝的决策树，其中，x 是输入向量，k 是独立同分布的随机向量，决定了单颗树的生长过程，速记森林的输出通常采用多数投票法得到。

由于样本数据的规模达到几亿甚至几十亿，单机版的随机森林已经不能处理海量规模的，通常使用并行版的随机森林。

假设样本数据的全集为 D，要训练 100 棵决策树，并行实现方案一般如下:

1、样本随机采样;

同时启动 100 个工作节点 worker，每个 worker 从 D 中随机采样出一个样本数据的子集 S，S 的大小一般远远小于 D，单台计算机可处理。

2、单个 worker 基于 S、应用 CART 算法训练决策树。

在训练决策树时，对于非连续特征，一般是计算该特征的基尼系数 Gini，基于最佳基尼系数 Gini 进行分裂。

在这种方案中，由于每个 worker 都是从样本数据的全集中采样子集，

因此，需要扫面一次样本数据的全集，数据读取量大，耗费较多的时间进行读取，使得模型的迭代更新时间较长、训练效率较低。

在计算基尼系数 Gini 时，通常需要使用穷举法，即假设有 n 个特征，且 CART 树是二分类的，则所有分支的组合有（2^{n-1}-1）种，需要计算（2^{n-1}-1）次基尼系数 Gini，复杂度为 $O(2^{n-1}-1)$，计算的复杂度为指数级别，在训练决策树时耗费大量的时间，同样使得使得模型的迭代更新时间较长、训练效率较低。

发明内容

鉴于上述问题，提出了本申请实施例以便提供一种克服上述问题或者至少部分地解决上述问题的一种基于随机森林的模型训练方法和相应的一种基于随机森林的模型训练装置。

为了解决上述问题，本申请实施例公开了一种基于随机森林的模型训练方法，包括:

将工作节点划分成一个或多个分组;

由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据;

由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象。

优选地，每个分组中的工作节点包括一个或多个第一工作节点以及一个或多个第二工作节点;

所述由每个分组中的工作节点从预置的样本数据中进行随机采样的，获得目标样本数据步骤包括:

在每个分组中，由每个第一工作节点从预置的样本数据中读取部分样本数据;

由每个第一工作节点将读取的部分样本数据随机分发至每个第二工作节点中，以分发至第二工作节点的样本数据作为目标样本数据。

优选地，所述由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象的步骤包括：

在每个分组中，由每个第二工作节点采用所述目标样本数据训练一个决策树对象。

优选地，所述由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象的步骤包括：

当所述目标样本数据的属性信息的值为枚举值时，计算所述属性信息的值的权重；

按照所述权重对所述属性信息的值进行排序；

采用排序后的属性信息的值计算基尼系数；

按照所述基尼系数针对决策树对象的树节点进行分裂处理。

优选地，所述计算所述属性信息的值的权重的步骤包括：

当所述属性信息的分类列为二分类时，计算所述属性信息的值对于所述分类列的频率；

对所述频率进行归一化，获得权重。

优选地，所述计算所述属性信息的值的权重的步骤包括：

当所述属性信息的分类列为多分类时，计算所述属性信息的值针对所述分类列的权重概率矩阵，其中，所述权重概率矩阵的横坐标为所述属性信息的值、纵坐标为所述分类列的值；

对所述权重概率矩阵进行主成分分析，获得最大特征值对应的特征向量；

将所述权重概率矩阵乘以所述特征向量，获得权重。

优选地，所述采用排序后的属性信息的值计算基尼系数的步骤包括：

按照排序的顺序依次将排序后的属性信息的值划分为两个属性子集；

依次采用所述两个属性子集计算基尼系数。

本申请实施例还公开了一种基于随机森林的模型训练装置，包括：

分组划分模块，用于将工作节点划分成一个或多个分组；

随机采样模块，用于由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据；

决策树训练模块，用于由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象。

优选地，每个分组中的工作节点包括一个或多个第一工作节点以及一个或多个第二工作节点；

所述随机采样模块包括：

部分数据读取子模块，用于在每个分组中，由每个第一工作节点从预置的样本数据中读取部分样本数据；

数据随机分发子模块，用于由每个第一工作节点将读取的部分样本数据随机分发至每个第二工作节点中，以分发至第二工作节点的样本数据作为目标样本数据。

优选地，所述决策树训练模块包括：

节点训练子模块，用于在每个分组中，由每个第二工作节点采用所述目标样本数据训练一个决策树对象。

优选地，所述决策树训练模块包括：

权重计算子模块，用于在所述目标样本数据的属性信息的值为枚举值时，计算所述属性信息的值的权重；

排序子模块，用于按照所述权重对所述属性信息的值进行排序；

基尼系数计算子模块，用于采用排序后的属性信息的值计算基尼系数；

分裂子模块，用于按照所述基尼系数针对决策树对象的树节点进行分裂处理。

优选地，所述权重计算子模块包括:

频率计算单元，用于在所述属性信息的分类列为二分类时，计算所述属性信息的值对于所述分类列的频率;

归一化单元，用于对所述频率进行归一化，获得权重。

优选地，所述权重计算子模块包括:

权重概率矩阵计算单元，用于在所述属性信息的分类列为多分类时，计算所述属性信息的值针对所述分类列的权重概率矩阵，其中，所述权重概率矩阵的横坐标为所述属性信息的值、纵坐标为所述分类列的值;

主成分分析单元，用于对所述权重概率矩阵进行主成分分析，获得最大特征值对应的特征向量;

权重获得单元，用于将所述权重概率矩阵乘以所述特征向量，获得权重。

优选地，所述基尼系数计算子模块包括:

子集划分单元，用于按照排序的顺序依次将排序后的属性信息的值划分为两个属性子集;

子集计算单元，用于依次采用所述两个属性子集计算基尼系数。

本申请实施例包括以下优点:

本申请实施例将工作节点划分成一个或多个分组，由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据，进而训练决策树对象，因此，每个分组中的工作节点仅是读取部分的样本数据，而不需要扫描一次完整的样本数据，大大降低了数据的读取量，减少了耗费的时间，进而减少模型的迭代更新时间、提高训练效率。

本申请实施例对此枚举值的属性，通过重要性排序的方式计算分裂

点，无需进行穷举，大大减少了分裂点的计算量，假设属性有 n 个值，通过重要性排序的方式计算分裂点的方式，计算的复杂度可以从穷举法的 $O(2^{n-1}-1)$，降低到 O(n)，减少了训练时间的耗费，进而减少模型的迭代更新时间、提高训练效率。

附图说明

图 1 是本申请的一种基于随机森林的模型训练方法实施例的步骤流程图；

图 2 是本申请实施例的一种分组示例图；

图 3 是本申请实施例的一种在 Hadoop 的分组中进行模型训练的流程示例图；

图 4 是本申请的一种基于随机森林的模型训练装置实施例的结构框图。

具体实施方式

为使本申请的上述目的、特征和优点能够更加明显易懂，下面结合附图和具体实施方式对本申请作进一步详细的说明。

参照图 1，示出了本申请的一种基于随机森林的模型训练方法实施例的步骤流程图，具体可以包括如下步骤：

步骤 101，将工作节点划分成一个或多个分组；

在本申请实施例中，工作节点可以为训练模型的计算节点，可以部署在单台计算机中，也可以应用在计算机集群中，如分布式系统，本申请实施例对此不加以限制。

对于单台计算机而言，工作节点（worker）可以为 CPU（Central

Processing Unit，中央处理器）的内核（Core），对于计算机集群，工作节点可以为单台计算机。

在本申请实施例中，可以按照样本数据的数据量、决策树的数量等因素，如图 2 所示，将工作节点划分为一个或多个分组（虚线框部分），每个分组中的工作节点包括一个或多个第一工作节点以及一个或多个第二工作节点。

其中，每个分组负责处理一份完整的样本数据，组内第一工作节点随机分发样本数据至第二工作节点，第二工作节点采用分发的样本数据训练决策树。

一般而言，考虑了系统的承受能力以及运算速度，分组的数目与决策树对象的数量成正比，例如，分组的数目=决策树的数量/100。

单个分组内，第一工作节点的数量与样本数据的数据量成正比，一个第二工作节点训练一棵决策树。

为使本领域技术人员更好地理解本申请实施例，在本申请实施例中，将 Hadoop 作为计算机集群的一种实施例进行说明。

Hadoop 主要包括两部分，一是分布式文件系统（Hadoop Distributed File System，HDFS），另外是分布式计算框架，即 MapReduce。

HDFS 是一个高度容错性的系统，能提供高吞吐量的数据访问，适合那些有着超大数据集（large data set）的应用程序。

MapReduce 是一套从海量源数据提取分析元素最后返回结果集的编程模型，其基本原理可以是将大的数据分析分成小块逐个分析，最后再将提取出来的数据汇总分析。

在 Hadoop 中，用于执行 MapReduce 的机器角色有两个：一个是 JobTracker，另一个是 TaskTracker。JobTracker 可以用于调度工作，TaskTracker 可以用于执行工作。

进一步而言，在 Hadoop 中 TaskTracker 可以指所述分布式系统的处理节点，该处理节点可以包括一个或多个映射（Map）节点和一个或多个化简（Reduce）节点。

在分布式计算中，MapReduce 负责处理了并行编程中分布式存储、工作调度、负载均衡、容错均衡、容错处理以及网络通信等复杂问题，把处理过程高度抽象为两个函数：映射函数（map 函数）和规约函数（reduce 函数），map 函数可以把任务分解成多个任务，reduce 函数可以把分解后的多任务处理的结果汇总起来。

在 Hadoop 中，每个 MapReduce 的任务可以被初始化为一个 Job，每个 Job 又可以分为两种阶段：map 阶段和 reduce 阶段。这两个阶段分别用两个函数表示，即 map 函数和 reduce 函数。

map 函数可以接收一个<key,value>形式的输入（Input），然后同样产生一个<key,value>形式的中间输出（Output），Hadoop 函数可以接收一个如<key,(list of values)>形式的输入（Input），然后对这个 value 集合进行处理，每个 reduce 函数产生 0 或 1 个输出（Output），reduce 函数的输出也是<key,value>形式的。

对于分组而言，第一工作节点可以为 Map 节点，第二工作节点可以为 Raduce 节点。

步骤 102，由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据；

在具体实现中，在每个分组中，读取预置的样本数据，即样本数据的全集，可以由每个第一工作节点从预置的样本数据中读取部分样本数据，即样本数据的子集。

由每个第一工作节点将读取的部分样本数据随机分发至每个第二工作节点中，以分发至第二工作节点的样本数据作为目标样本数据。

对于每条样本数据，第一工作节点均读取一次，但是否会分发到第二工作节点中是不确定的，即随机分发（采样）。

例如，如图 2 所示，某一条样本数据由第一工作节点 A1 读取，针对第二工作节点 B1、B2、B3、B4、B5，分别生成一随机值，如果该随机值大于 0.5，则分发到该第二工作节点中，反之，则不分发到该第二工作节点，对于该条样本数据，分发随机了 5（第二工作节点的数量）次。

同理，对于第一工作节点 A2、A3 读取的样本数据，也可以随机分发至第二工作节点 B1、B2、B3、B4、B5。

如图 3 所示，在 Hadoop 中，一个分组的 Map 节点和 Raduce 节点处理一份完整的样本数据，每个 Map 节点读取部分样本数据，随机分发至 Raduce 节点中。

即在 Map 节点中，可以定义 map 函数为随机分发，以将 Map 节点的样本数据分发到 Reduce 节点中。

Map 节点从输入的样本数据中抽取出键值对，每一个键值对都作为参数传递给 map 函数，map 函数产生的中间键值对被缓存在内存中。

Map 节点中的 map 函数的输出经由 MapReduce 框架处理后，最后分发到 Reduce 节点中的 reduce 函数。

步骤 103，由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象。

每条样本数据，通常包括一个样本对象、一个或多个属性信息、一个分类标签。

对于随机采样之后的目标样本数据为一个数据集合，一般为二维数组的形式，即包括一组样本对象、一组或多组属性信息、一组分类标签（又称分类列）。

一个目标样本数据的示例如下表所示：

表 1

样本对象	体温	表面覆盖	胎生	产蛋	能飞	水生	有腿	冬眠	分类列
人	恒温	毛发	是	否	否	否	是	否	哺乳类
巨蟒	冷血	鳞片	否	是	否	否	否	是	爬行类
鲑鱼	冷血	鳞片	否	是	否	是	否	否	鱼类
鲸	恒温	毛发	是	否	否	是	否	否	哺乳类
蛙	冷血	无	否	是	否	有时	是	是	两栖类
巨蜥	冷血	鳞片	否	是	否	否	是	否	爬行类
蝙蝠	恒温	毛发	是	否	是	否	是	否	哺乳类
猫	恒温	皮	是	否	否	否	是	否	哺乳类
豹纹鲨	冷血	鳞片	是	否	否	是	否	否	鱼类
海龟	冷血	鳞片	否	是	否	有时	是	否	爬行类
豪猪	恒温	刚毛	是	否	否	否	是	是	哺乳类
鳗	冷血	鳞片	否	是	否	是	否	否	鱼类
蝰蛇	冷血	无	否	是	否	有时	是	是	两栖类

其中，属性信息包括体温、表面覆盖、胎生、产蛋、能飞、水生、有腿、冬眠。

决策树（对象）是一种由节点和有向边构成的树状结构，训练时，在每一个非叶子节点针对某一属性进行分类。

在具体实现中，在每个分组中，由每个第二工作节点采用所述目标样本数据训练一个决策树对象。

如图 3 所示，在 Hadoop 中，若 Map 节点读取的部分样本数据随机分发至 Raduce 节点中，则 Raduce 节点可以采用该分发的样本数据（即目标样本数据）训练决策树。

在训练决策树时，在每一个非叶子节点针对某一属性进行分裂、迭代这一过程，直到每个叶子节点上的样本均处于单一的类别或者每个属性都被选择过为止。叶子节点代表分类的结果，从根节点到叶子节点的完整路径代表一种决策过程，决策树的训练本质是节点如何进行分裂。

训练得到的决策树一般是二叉树，少数情况下也存在非二叉树的情况，具体的训练过程如下:

（1）、构造决策树的根节点，为全体目标训练样本数据的集合 T;

（2）、通过计算信息增益或基尼系数选择出 T 中区分度最高的属性，分割形成左子节点和右子节点;

（3）、在剩余的属性空间中，针对每一个子节点的样本数据，重复步骤 2 的过程，若满足以下条件之一则标记为叶子节点，此节点分裂结束:

a、该节点上所有样本数据都属于同一个分类;

b、没有剩余的属性可用以分裂;

c、当前数据集的样本数据个数小于某个给定的值;

d、决策树的深度大于设定的值。

本申请实施例将工作节点划分成一个或多个分组，由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据，进而训练决策树对象，因此，每个分组中的工作节点仅是读取部分的样本数据，而不需要扫描一次完整的样本数据，大大降低了数据的读取量，减少了耗费的时间，进而减少模型的迭代更新时间、提高训练效率。

在本申请的一个实施例中，步骤 103 可以包括如下子步骤:

子步骤 S11，当所述目标样本数据的属性信息的值为枚举值时，计算所述属性信息的值的权重;

在实现应用中，属性信息的值一般分为连续值和枚举值，枚举值又称离散值，即不连续的值。

例如，表 1 中体温的值为冷血、恒温，属于枚举值。

在本申请实施例中，针对枚举值的属性信息，利用其重要性（权重）排序来计算最佳分裂点，来提升加速比。

在一个示例中，当属性信息的分类列为二分类（即具有两个分类）时，计算该属性信息的值对于分类列的频率，对频率进行归一化，获得权重。

在另一个示例中，当属性信息的分类列为多分类（即具有三个或三个以上的分类）时，计算属性信息的值针对分类列的权重概率矩阵，其中，权重概率矩阵的横坐标为属性信息的值、纵坐标为分类列的值。

对所述权重概率矩阵进行主成分分析（Principal Component Analysis , PCA），获得最大特征值对应的特征向量，将权重概率矩阵乘以特征向量，获得权重。

子步骤 S12，按照所述权重对所述属性信息的值进行排序;

在具体实现中，可以按照权重对属性信息的值进行顺序排序，也可以倒序排序，本申请实施例对此不加以限制。

子步骤 S13，采用排序后的属性信息的值计算基尼系数;

基尼系数 Gini，可以用于决策树的节点的分裂标准，样本总体内包含的类别越杂乱，Gini 指数就越大。

在实际应用中，可以按照排序的顺序依次将排序后的属性信息的值划分为两个属性子集。

假设按权重排序得到的有序属性信息的值序列为 f =(a1, a2,

a3……an)，那么，可以划分为左子树（属性子集）为 a_1~a_i，右子树（属性子集）为 a_{i+1}~a_n，其中，i =1，2，…，n-1。

依次采用两个属性子集计算基尼系数 Gini。

假设有 k 个分类，样本数据属于第 i 类的概率为 p_i，则基尼指数 Gini 定义为：

$$Gini = \sum_{i=1}^{k} p_i(1-p_i) = 1-\sum_{i=1}^{k} p_i^2$$

如果数据集合 D 被划分成 D_1 和 D_2 两部分，则在该条件下，集合 D 的基尼增益定义为：

$$Gini(D) = \frac{|D_1|}{|D_2|}Gini(D_1) + \frac{|D_2|}{|D_1|}Gini(D_2)$$

子步骤 S14，按照所述基尼系数针对决策树对象的树节点进行分裂处理。

基尼指数 Gini 表示数据集合的不确定性，基尼指数 Gini 的值越大，样本属于某个分类的不确定性也就越大。因此，最好的选取特征划分就是使得数据集合的基尼指数 Gini 最小的划分。

本申请实施例对此枚举值的属性，通过重要性排序的方式计算分裂点，无需进行穷举，大大减少了分裂点的计算量，假设属性有 n 个值，通过重要性排序的方式计算分裂点的方式，计算的复杂度可以从穷举法的 $O(2^{n-1}-1)$，降低到 O(n)，减少了训练时间的耗费，进而减少模型的迭代更新时间、提高训练效率。

需要说明的是，对于方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请实施例并不受所描述的动作顺序的限制，因为依据本申请实施例，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述

的实施例均属于优选实施例，所涉及的动作并不一定是本申请实施例所必须的。

参照图 4，示出了本申请的一种基于随机森林的模型训练装置实施例的结构框图，具体可以包括如下模块：

分组划分模块 401，用于将工作节点划分成一个或多个分组；

随机采样模块 402，用于由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据；

决策树训练模块 403，用于由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象。

在本申请的一个实施例中，每个分组中的工作节点包括一个或多个第一工作节点以及一个或多个第二工作节点；

所述随机采样模块 401 可以包括如下子模块：

部分数据读取子模块，用于在每个分组中，由每个第一工作节点从预置的样本数据中读取部分样本数据；

数据随机分发子模块，用于由每个第一工作节点将读取的部分样本数据随机分发至每个第二工作节点中，以分发至第二工作节点的样本数据作为目标样本数据。

在本申请的一个实施例中，所述决策树训练模块 403 可以包括如下子模块：

节点训练子模块，用于在每个分组中，由每个第二工作节点采用所述目标样本数据训练一个决策树对象。

在本申请的一个实施例中，所述决策树训练模块 403 可以包括如下子模块：

权重计算子模块，用于在所述目标样本数据的属性信息的值为枚举值时，计算所述属性信息的值的权重；

排序子模块，用于按照所述权重对所述属性信息的值进行排序；

基尼系数计算子模块，用于采用排序后的属性信息的值计算基尼系数；

分裂子模块，用于按照所述基尼系数针对决策树对象的树节点进行分裂处理。

在本申请的一个实施例中，所述权重计算子模块可以包括如下单元：

频率计算单元，用于在所述属性信息的分类列为二分类时，计算所述属性信息的值对于所述分类列的频率；

归一化单元，用于对所述频率进行归一化，获得权重。

在本申请的一个实施例中，所述权重计算子模块可以包括如下单元：

权重概率矩阵计算单元，用于在所述属性信息的分类列为多分类时，计算所述属性信息的值针对所述分类列的权重概率矩阵，其中，所述权重概率矩阵的横坐标为所述属性信息的值、纵坐标为所述分类列的值；

主成分分析单元，用于对所述权重概率矩阵进行主成分分析，获得最大特征值对应的特征向量；

权重获得单元，用于将所述权重概率矩阵乘以所述特征向量，获得权重。

在本申请的一个实施例中，所述基尼系数计算子模块可以包括如下单元：

子集划分单元，用于按照排序的顺序依次将排序后的属性信息的值划分为两个属性子集；

子集计算单元，用于依次采用所述两个属性子集计算基尼系数。

对于装置实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

本领域内的技术人员应明白，本申请实施例的实施例可提供为方法、装置、或计算机程序产品。因此，本申请实施例可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请实施例可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

在一个典型的配置中，所述计算机设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM) 和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储

器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括非持续性的电脑可读媒体(transitory media)，如调制的数据信号和载波。

本申请实施例是参照根据本申请实施例的方法、终端设备(系统)、和计算机程序产品的流程图和／或方框图来描述的。应理解可由计算机程序指令实现流程图和／或方框图中的每一流程和／或方框、以及流程图和／或方框图中的流程和／或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理终端设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和／或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和／或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上，使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和／或方框图一个方框或多个方框中指定的功能的步骤。

尽管已描述了本申请实施例的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本申请实施例范围的所有变更和修改。

最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者终端设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者终端设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者终端设备中还存在另外的相同要素。

以上对本申请所提供的一种基于随机森林的模型方法和一种基于随机森林的模型装置，进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请

的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，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申请的限制。

权 利 要 求 书

1、一种基于随机森林的模型训练方法，其特征在于，包括:

将工作节点划分成一个或多个分组;

由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据;

由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象。

2、根据权利要求 1 所述的方法，其特征在于，每个分组中的工作节点包括一个或多个第一工作节点以及一个或多个第二工作节点;

所述由每个分组中的工作节点从预置的样本数据中进行随机采样的，获得目标样本数据步骤包括:

在每个分组中，由每个第一工作节点从预置的样本数据中读取部分样本数据;

由每个第一工作节点将读取的部分样本数据随机分发至每个第二工作节点中，以分发至第二工作节点的样本数据作为目标样本数据。

3、根据权利要求 2 所述的方法，其特征在于，所述由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象的步骤包括:

在每个分组中，由每个第二工作节点采用所述目标样本数据训练一个决策树对象。

4、根据权利要求 1 或 2 或 3 所述的方法，其特征在于，所述由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象的步骤包括:

当所述目标样本数据的属性信息的值为枚举值时，计算所述属性信息的值的权重;

按照所述权重对所述属性信息的值进行排序;

采用排序后的属性信息的值计算基尼系数;

按照所述基尼系数针对决策树对象的树节点进行分裂处理。

5、根据权利要求 4 所述的方法，其特征在于，所述计算所述属性信息的值的权重的步骤包括:

当所述属性信息的分类列为二分类时，计算所述属性信息的值对于所述分类列的频率;

对所述频率进行归一化，获得权重。

6、根据权利要求 4 所述的方法，其特征在于，所述计算所述属性信息的值的权重的步骤包括:

当所述属性信息的分类列为多分类时，计算所述属性信息的值针对所述分类列的权重概率矩阵，其中，所述权重概率矩阵的横坐标为所述属性信息的值、纵坐标为所述分类列的值;

对所述权重概率矩阵进行主成分分析，获得最大特征值对应的特征向量;

将所述权重概率矩阵乘以所述特征向量，获得权重。

7、根据权利要求 4 所述的方法，其特征在于，所述采用排序后的属性信息的值计算基尼系数的步骤包括:

按照排序的顺序依次将排序后的属性信息的值划分为两个属性子集;

依次采用所述两个属性子集计算基尼系数。

8、一种基于随机森林的模型训练装置，其特征在于，包括:

分组划分模块，用于将工作节点划分成一个或多个分组;

随机采样模块，用于由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据;

决策树训练模块，用于由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象。

9、根据权利要求 8 所述的装置，其特征在于，每个分组中的工作节点包括一个或多个第一工作节点以及一个或多个第二工作节点；

所述随机采样模块包括：

部分数据读取子模块，用于在每个分组中，由每个第一工作节点从预置的样本数据中读取部分样本数据；

数据随机分发子模块，用于由每个第一工作节点将读取的部分样本数据随机分发至每个第二工作节点中，以分发至第二工作节点的样本数据作为目标样本数据。

10、根据权利要求 9 所述的装置，其特征在于，所述决策树训练模块包括：

节点训练子模块，用于在每个分组中，由每个第二工作节点采用所述目标样本数据训练一个决策树对象。

11、根据权利要求 8 或 9 或 10 所述的装置，其特征在于，所述决策树训练模块包括：

权重计算子模块，用于在所述目标样本数据的属性信息的值为枚举值时，计算所述属性信息的值的权重；

排序子模块，用于按照所述权重对所述属性信息的值进行排序；

基尼系数计算子模块，用于采用排序后的属性信息的值计算基尼系数；

分裂子模块，用于按照所述基尼系数针对决策树对象的树节点进行分裂处理。

12、根据权利要求 11 所述的装置，其特征在于，所述权重计算子模块包括：

频率计算单元，用于在所述属性信息的分类列为二分类时，计算所述属性信息的值对于所述分类列的频率；

归一化单元，用于对所述频率进行归一化，获得权重。

13、根据权利要求 11 所述的装置，其特征在于，所述权重计算子模块包括:

权重概率矩阵计算单元，用于在所述属性信息的分类列为多分类时，计算所述属性信息的值针对所述分类列的权重概率矩阵，其中，所述权重概率矩阵的横坐标为所述属性信息的值、纵坐标为所述分类列的值;

主成分分析单元，用于对所述权重概率矩阵进行主成分分析，获得最大特征值对应的特征向量;

权重获得单元，用于将所述权重概率矩阵乘以所述特征向量，获得权重。

14、根据权利要求 11 所述的装置，其特征在于，所述基尼系数计算子模块包括:

子集划分单元，用于按照排序的顺序依次将排序后的属性信息的值划分为两个属性子集;

子集计算单元，用于依次采用所述两个属性子集计算基尼系数。

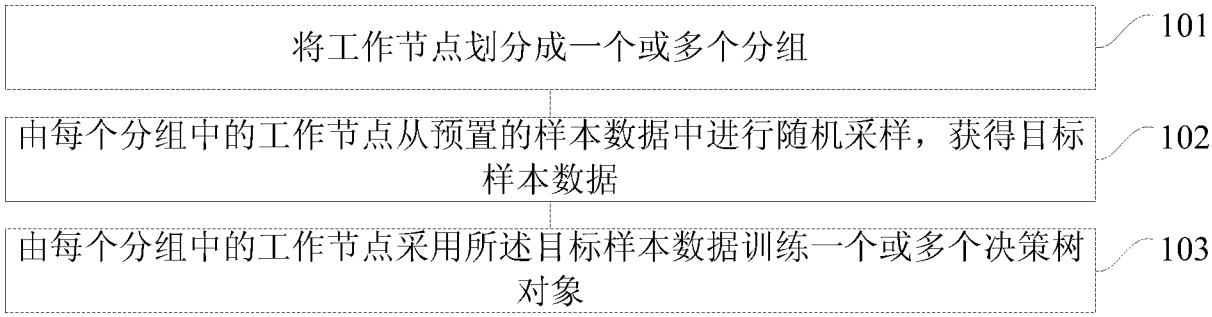

将工作节点划分成一个或多个分组
101
由每个分组中的工作节点从预置的样本数据中进行随机采样，获得目标样本数据
102
由每个分组中的工作节点采用所述目标样本数据训练一个或多个决策树对象
103

图 1

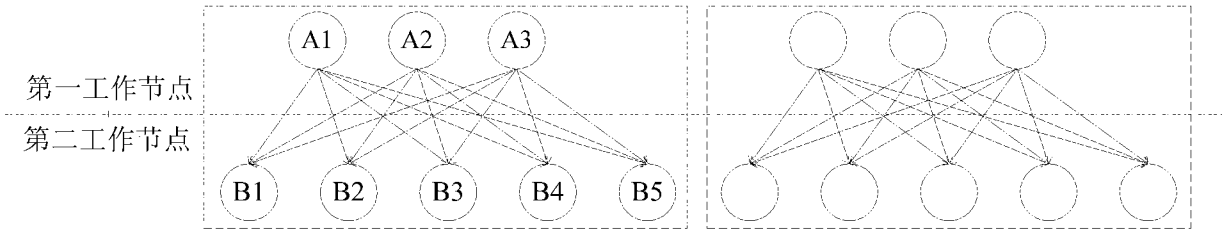

第一工作节点
第二工作节点
A1
A2
A3
B1
B2
B3
B4
B5

图 2

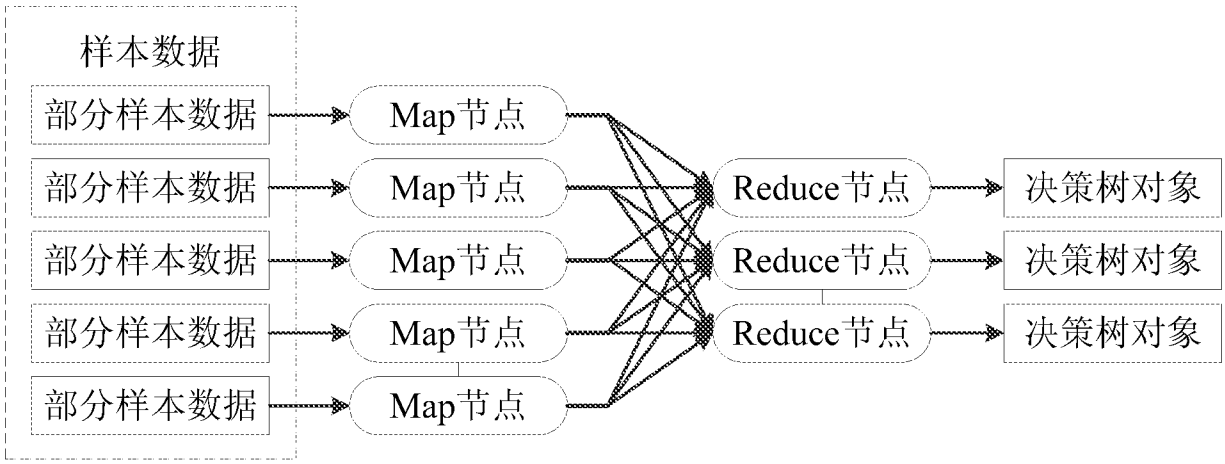

样本数据
部分样本数据
部分样本数据
部分样本数据
部分样本数据
部分样本数据
Map节点
Map节点
Map节点
Map节点
Map节点
Reduce节点
Reduce节点
Reduce节点
决策树对象
决策树对象
决策树对象

图 3

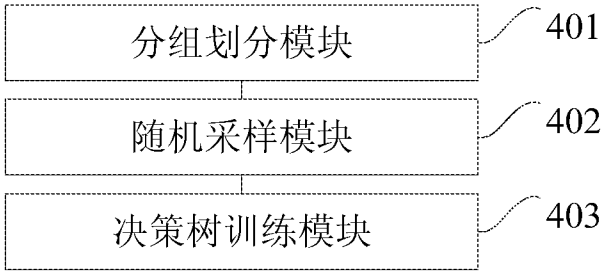

分组划分模块
随机采样模块
决策树训练模块
401
402
403

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/077704

A. CLASSIFICATION OF SUBJECT MATTER

G06F 21/62 (2013.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI: node, division of labor, fragmentation, random forest, model, train+, sampl+, worker, group+, gini, decision tree, map, reduce

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 104392250 A (LANGCHAO ELECTRONIC INFORMATION INDUSTRY CO., LTD.) 04 March 2015 (04.03.2015) claims 1-6	1-14
Y	LIU, Kai et al. 'Realization of RBM Training Based on Hadoop-GPU' Microelectronics & Computer, vol. 32, no. 9, 30 September 2015 (30.09.2015) , 1SSN:1000-7180, pages 70-72	1-14
A	CN 103473231 A (SHENZHEN INSTITUTE OF ADVANCED TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 25 December 2013 (25.12.2013) the whole document	1-14
A	CN 105303262 A (HOHAI UNIVERSITY) 03 February 2016 (03.02.2016) the whole document	1-14
A	JP 2006236153 A (DAINIPPON PHARM CO., LTD. et al.) 07 September 2006 (07.09.2006) the whole document	1-14

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
08 June 2017	30 June 2017
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer LIN, Wanjuan Telephone No. (86-10) 82246854

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/077704

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104392250 A	04 March 2015	None	
CN 103473231 A	25 December 2013	None	
CN 105303262 A	03 February 2016	None	
JP 2006236153 A	07 September 2006	None	

国际检索报告

国际申请号
PCT/CN2017/077704

A. 主题的分类

G06F 21/62(2013.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F,G06K

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNPAT,WPI,EPODOC,CNKI:随机森林，模型，训练，样本，采样，取样，节点，分组，分工，分片，基尼，决策树，random forest, model, train+, sampl+, worker, group+, gini,decision tree, map, reduce

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	CN 104392250 A (浪潮电子信息产业股份有限公司) 2015年 3月 4日 (2015-03-04) 权利要求1-6	1-14
Y	刘凯 等. "基于Hadoop-GPU的RBM云计算实现" 微电子学与计算机, 第32卷，第9期，2015年 9月 30日 (2015-09-30)， ISSN：1000-7180, 第70-72页	1-14
A	CN 103473231 A (深圳先进技术研究院) 2013年 12月 25日 (2013-12-25) 全文	1-14
A	CN 105303262 A (河海大学) 2016年 2月 3日 (2016-02-03) 全文	1-14
A	JP 2006236153 A (DAINIPPON PHARM CO., LTD. 等) 2006年 9月 7日 (2006-09-07) 全文	1-14

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2017年 6月 8日	2017年 6月 30日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 林婉娟 电话号码 (86-10)82246854

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/077704

检索报告引用的专利文件	公布日（年/月/日）	同族专利	公布日（年/月/日）
CN 104392250 A	2015年 3月 4日	无	
CN 103473231 A	2013年 12月 25日	无	
CN 105303262 A	2016年 2月 3日	无	
JP 2006236153 A	2006年 9月 7日	无	