

(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 8 月 3 日 (03.08.2017)



(10) 国际公布号
WO 2017/129030 A1

- (51) 国际专利分类号:
G06F 11/22 (2006.01)
- (21) 国际申请号: PCT/CN2017/071695
- (22) 国际申请日: 2017 年 1 月 19 日 (19.01.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610065807.2 2016 年 1 月 29 日 (29.01.2016) 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): 丁永明 (DING, Yongming) [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)。周俊 (ZHOU, Jun) [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)。崔卿 (CUI, Qing) [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)。瞿神全 (QU, Shenquan) [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: DISK FAILURE PREDICTION METHOD AND APPARATUS

(54) 发明名称: 磁盘的故障预测方法和装置

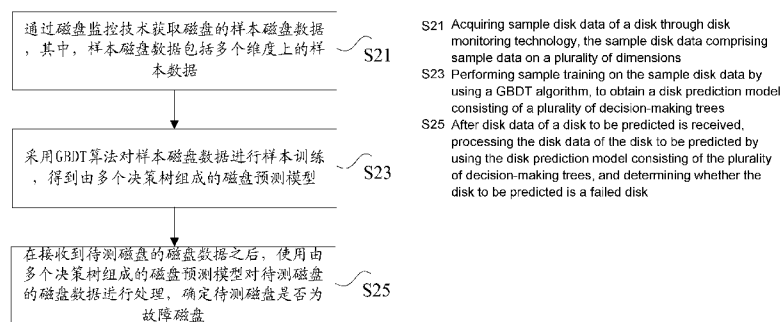


图 2

(57) Abstract: Disclosed are a disk failure prediction method and apparatus. The method comprises: acquiring sample disk data of a disk through disk monitoring technology, the sample disk data comprising sample data on a plurality of dimensions (S21); performing sample training on the sample disk data by using a GBDT algorithm, to obtain a disk prediction model consisting of a plurality of decision-making trees (S23); and after disk data of a disk to be predicted is received, processing the disk data of the disk to be predicted by using the disk prediction model consisting of the plurality of decision-making trees, and determining whether the disk to be predicted is a failed disk (S25). The method solves the technical problem in the prior art of an inaccurate prediction result caused by the fact that some factors resulting in hard disk failures cannot be collected or quantized in a hard disk failure prediction system.

(57) 摘要: 一种磁盘的故障预测方法和装置。其中, 该方法包括: 通过磁盘监控技术获取磁盘的样本磁盘数据, 其中, 样本磁盘数据包括多个维度上的样本数据 (S21); 采用 GBDT 算法对样本磁盘数据进行样本训练, 得到由多个决策树组成的磁盘预测模型 (S23); 在接收到待测磁盘的磁盘数据之后, 使用由多个决策树组成的磁盘预测模型对待测磁盘的磁盘数据进行处理, 确定待测磁盘是否为故障磁盘 (S25)。该方法解决了现有技术中的一些容易致使硬盘故障的因素不能被采集或量化导致的预测结果不准确的技术问题。



WO 2017/129030 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))。

磁盘的故障预测方法和装置

技术领域

本发明涉及磁盘领域，具体而言，涉及一种磁盘的故障预测方法和装置。

背景技术

目前，硬盘是存储数据的主要介质，硬盘一旦出故障，便会造成巨大的数据损失。因此如何保证硬盘的稳定性能非常重要。在通常状态下，硬盘在 24 小时中出错的概率在是万分之一左右，当一台服务器具有十块硬盘时，服务器硬盘出错的概率就会上升到千分之一，而随着当前网站等业务的发展，服务器需要使用的硬盘会越来越多，多块硬盘同时出错的概率也会提升。

通常情况下，数据存储通常会有多个备份，如 mysql 主备库，GFS 文件默认 3 个备份。在大量数据存储平台上，如果多个硬盘同时出故障，那么这些硬盘上存储着同一个文件的备份的概率就会很高，即如果多块硬盘同时出现故障，就会导致一些文件的丢失，对于一些线上的服务，大都依赖于服务器中存储的海量数据，如果硬盘出故障，就会导致上述在线服务异常，甚至暂停使用。

由于上述原因，需要具有预测硬盘是否会出错的系统需要有一套系统能提前告诉我们哪些硬盘会出错，数据可能丢失导致硬盘故障的原因有很多，最常见的有以下几种：外部振动、温度和湿度、电器元件损坏、声音和灰尘，在上述因素中，有些因素能够被采集到，比如温度和湿度、一些元器件数据，但是更多的数据无法被采集和量化，因此便会导致预测结果不准确。

针对现有技术的硬盘故障预测系统中一些容易致使硬盘故障的因素

不能被采集或量化导致的预测结果不准确的问题，目前尚未提出有效的解决方案。

发明内容

5 本发明实施例提供了一种磁盘的故障预测方法和装置，以至少解决现有技术的硬盘故障预测系统中一些容易致使硬盘故障的因素不能被采集或量化导致的预测结果不准确的技术问题。

根据本发明实施例的一个方面，提供了一种磁盘的故障预测方法，包括：通过磁盘监控技术获取磁盘的样本磁盘数据，其中，样本磁盘数据包
10 括多个维度上的样本数据；采用 GBDT 算法对样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型；在接收到待测磁盘的磁盘数据之后，使用由多个决策树组成的磁盘预测模型对待测磁盘的磁盘数据进行处理，确定待测磁盘是否为故障磁盘。

根据本发明实施例的另一方面，还提供了一种磁盘的故障预测装置，
15 包括：通过磁盘监控技术获取磁盘的样本磁盘数据，其中，样本磁盘数据包括多个维度上的样本数据；采用 GBDT 算法对样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型；在接收到待测磁盘的磁盘数据之后，使用由多个决策树组成的磁盘预测模型对待测磁盘的磁盘数据进行处理，确定待测磁盘是否为故障磁盘。

20 在本发明实施例中，采用通过磁盘监控技术获取磁盘的样本磁盘数据，其中，样本磁盘数据包括多个维度上的样本数据；采用 GBDT 算法对样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型方式，通过在接收到待测磁盘的磁盘数据之后，使用由多个决策树组成的磁盘预测模型对待测磁盘的磁盘数据进行处理，达到了确定待测磁盘是否为故障磁
25 盘的目的，从而实现了预测磁盘故障状态的技术效果，进而解决了现有技术的硬盘故障预测系统中一些容易致使硬盘故障的因素不能被采集或量化导致的预测结果不准确的技术问题。

附图说明

此处所说明的附图用来提供对本发明的进一步理解，构成本申请的一部分，本发明的示意性实施例及其说明用于解释本发明，并不构成对本发明的不当限定。在附图中：

图 1 是根据本发明实施例的一种磁盘的故障预测方法的计算机终端的硬件结构框图；

图 2 是根据本发明实施例的一种磁盘的故障预测方法的流程图；

图 3 是根据本发明实施例的一种使用 GBDT 算法对样本磁盘数据进行训练的示意图；

图 4 为根据本发明实施例的一种使用 GBDT 算法计算磁盘预测值的示意图；

图 5 是根据本发明实施例的一种可选的磁盘的故障预测方法的流程图；

图 6 是根据本发明实施例的一种磁盘的故障预测装置的结构示意图；

图 7 是根据本发明实施例的一种可选的磁盘的故障预测装置的结构示意图；

图 8 是根据本发明实施例的一种可选的磁盘的故障预测装置的结构示意图；

图 9 是根据本发明实施例的一种可选的磁盘的故障预测装置的结构示意图；

图 10 是根据本发明实施例的一种可选的磁盘的故障预测装置的结构示意图；以及

图 11 是根据本发明实施例的一种计算机终端的结构框图。

具体实施方式

为了使本技术领域的人员更好地理解本发明方案，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分的实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本发明保护的范围。

需要说明的是，本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本发明的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外，术语“包括”和“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

实施例 1

根据本发明实施例，还提供了一种磁盘的故障预测方法实施例，需要说明的是，在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行，并且，虽然在流程图中示出了逻辑顺序，但是在某些情况下，可以以不同于此处的顺序执行所示出或描述的步骤。

本申请实施例一所提供的方法实施例可以在移动终端、计算机终端或者类似的运算装置中执行。以运行在计算机终端上为例，图 1 是根据本发明实施例的一种磁盘的故障预测方法的计算机终端的硬件结构框图。如图 1 所示，计算机终端 10 可以包括一个或多个（图中仅示出一个）处理器 102（处理器 102 可以包括但不限于微处理器 MCU 或可编程逻辑器件 FPGA 等的处理装置）、用于存储数据的存储器 104、以及用于通信功能的传输模块 106。本领域普通技术人员可以理解，图 1 所示的结构仅为示意，其并不对上述电子装置的结构造成限定。例如，计算机终端 10 还可包括比图 1 中所示更多或者更少的组件，或者具有与图 1 所示不同的配置。

存储器 104 可用于存储应用软件的程序以及模块，如本发明实施例中的磁盘的故障预测方法对应的程序指令/模块，处理器 102 通过运行存储在存储器 104 内的软件程序以及模块，从而执行各种功能应用以及数据处理，即实现上述的应用程序的漏洞检测方法。存储器 104 可包括高速随机存储器，还可包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器 104 可进一步包括相对于处理器 102 远程设置的存储器，这些远程存储器可以通过网络连接至计算机终端 10。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

传输装置 106 用于经由一个网络接收或者发送数据。上述的网络具体实例可包括计算机终端 10 的通信供应商提供的无线网络。在一个实例中，传输装置 106 包括一个网络适配器（Network Interface Controller，NIC），其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个实例中，传输装置 106 可以为射频（Radio Frequency，RF）模块，其用于通过无线方式与互联网进行通讯。

在上述运行环境下，本申请提供了如图 2 所示的一种磁盘的故障预测方法。图 2 是根据本发明实施例的一种磁盘的故障预测方法的流程图。

需要说明的是，对于前述的各方法实施例，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，本发明并不受所描述的动作顺序的限制，因为依据本发明，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，所涉及的动作和模块并不一定是本发明所必须的。

通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质（如 ROM/RAM、磁碟、光盘）中，包括若干指令用以使得一台终端设备（可

以是手机，计算机，服务器，或者网络设备等）执行本发明各个实施例所述的方法。

在上述运行环境下，本申请提供了如图 2 所示的反编译数据的处理方法。图 2 是根据本发明实施例一的反编译数据的处理方法的流程图，如图 2 所示，该方法包括：

步骤 21，通过磁盘监控技术获取磁盘的样本磁盘数据，其中，所述样本磁盘数据包括多个维度上的样本数据。

在上述步骤中，磁盘监控技术用于监测磁盘出厂后的使用过程中产生的各项磁盘数据，以预测磁盘的故障状态，使得磁盘使用者能够在磁盘发生故障之前便能知晓磁盘即将发生故障，从而对磁盘中的数据进行拷贝存储，避免数据的丢失。

在一种可选的实施例中，上述样本磁盘数据可以包括：底层数据读取错误率、启动/停止计数、重映射扇区数、通电时间累计、主轴起旋重试次数、磁盘校准重试次数、磁盘通电次数、温度以及写错误率，可以根据磁盘历史故障情况获取样本磁盘数据。例如，可以按照正负样本比例为 1:5 的比例进行样本获取，其中，正样本为存在故障的磁盘，负样本为不存在故障的磁盘。

此处需要说明的是，在通过磁盘监控技术获取磁盘的样本磁盘数据时，由于预测磁盘故障的各个机构使用的磁盘并不一定相同，且由于各个机构不同温湿度等环境因素对磁盘的影响，使得不同机构的磁盘的好坏比例并不相同，为了使样本磁盘数据的训练提供更可靠的样本磁盘数据，还可以根据机构的实际上磁盘损坏情况进行获取样本磁盘数据。

步骤 S23，采用 GBDT 算法对所述样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型。

在上述步骤中，GBDT(Gradient Boosting Decision Tree)为一种迭代的决策树算法，该算法由多棵决策树组成，并通过对所有决策树的结论进行累加，得到最终结果。上述决策树作为一种预测模型，是在上一层决策得

到的结果的基础上，进行下一层决策，包括决策点、状态结点、结果结点等参数，树中的每个节点表示被预测的对象，二每个分叉路径则代表该对象可能的属性。

在一种可选的实施例中，在上述样本磁盘为磁盘的 S.M.A.R.T 的原始值的情况下，对样本磁盘进行样本训练，例如，原始值大于等于预设原始值，可以认为该样本磁盘发生故障的概率较大，原始值小于预设值原始时，可以认为该样本磁盘发生故障的概率较小，因此在确定磁盘预测模型时，在样本磁盘的原始值大于等于预设原始值的情况下，确认该样本磁盘的属性为故障，在样本磁盘的原始值小于预设原始值的情况下，确认该样本磁盘的属性为非故障。建立具备上述决策能力的磁盘预测模型，即向决策树输入待检测磁盘时，若待检测磁盘的原始值大于等于预设原始值，决策树自动确认该待检测磁盘为故障的情况下，确认该样本磁盘的属性为故障，当样本磁盘的原始值小于预设原始值的情况下，确认该样本磁盘的属性为非故障。

步骤 S25，在接收到待测磁盘的磁盘数据之后，使用所述由多个决策树组成的磁盘预测模型对所述待测磁盘的磁盘数据进行处理，确定所述待测磁盘是否为故障磁盘。

在一种可选的实施例中，将样本磁盘的多个维度的值作为决策树的评价指标，得到多个决策树，再由多个决策树构成一个磁盘预测模型，对待检测磁盘进行检测。

此处值得注意的是，根据磁盘每一个维度得到的决策树可能相同，可能不相同，因此在使用多个决策树构成磁盘预测模型时，需要根据每个决策树在评价体系中的重要性，来确认每个决策树的权重值，从而得到磁盘预测模型。

此处需要说明的是，在通过磁盘监控技术获取磁盘的样本磁盘数据时，采用了磁盘检测技术，使得获取样本磁盘数据的过程更为简单，且获取的数据更为全面，为样本磁盘数据的训练提供了丰富的磁盘样本数据。在上述步骤中，采用 GBDT 算法对所述样本磁盘数据进行样本训练可以是分两

次或多次进行训练，以提高与训练结果对应的决策树构成的磁盘预测模型的准确率和召回率。

由此，本申请提供的上述实施例一的方案解决了现有技术的硬盘故障预测系统中一些容易致使硬盘故障的因素不能被采集或量化导致的预测结果不准确的技术问题。

根据本申请上述实施例，在一种优选的方案中，所述样本磁盘数据至少包括如下四个维度上的样本数据：原始值、标准值、最差值和累积值。

上述原始值为磁盘运行时的当前参数；上述标准值为正常磁盘运行时各项参数的数值；上述最差值为磁盘运行时，磁盘的各项检测参数曾出现过与正常值偏差最大的非正常值；上述累计值为磁盘的各项检测参数从磁盘使用至当前时刻的累计结果。

在一种可选的实施例中，磁盘的各项参数可以是对磁盘的各项属性进行描述的信息，可以包括错误读取率、加电次数、重新分配扇区数、旋转重试次数、磁盘校准重试次数以及奇偶校验错误率中的一项或多项，也可以包括磁盘的其他属性信息。

本申请上述步骤可以分别以上述四个维度上的样本数据得到多个不同的决策树。

在一种可选的实施例中，可以采用 HDTune、CrystalDiskInfo 等软件获取样本磁盘数据。

根据本申请上述实施例，在一种优选的方案中，在通过磁盘监控技术获取磁盘的样本磁盘数据之后，所述方法还包括：

步骤 S211，对所述每个维度上的样本数据进行如下任意一种或多种运算：差分运算、平方运算和分布求和运算，使得任意一个维度上的样本数据被扩展出新的维度上的样本数据。

在上述步骤中，对决策结果进行进一步运算，可将决策树根据运算结果拓展出新的维度，得到这一维度上的样本数据。

此处值得注意的时，每个维度的样本数据都可以进行多种运算以在这一维度的基础上得到更多维度的样本数据，在有四个维度的基础上，每个维度再分别进行差分运算、平方运算和分布求和运算，便能够得到十六个维度的样本数据，且通过每个维度的样本数据进行决策的侧重点均不同。

5 在一种可选的实施例，仍以原始值这一维度的样本数据为例，对原始值的样本数据进行差分运算、平方运算和分布求和运算，由此得到新的四个维度的样本数据，采用新的四个维度的样本数据最为决策指标进行训练，并得到新的四个决策树。

10 根据本申请上述实施例，在一种优选的方案中，采用 GBDT 算法对所述样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型，包括：

步骤 S231，以所有磁盘的样本磁盘数据作为训练数据，并采用默认值初始化所述训练数据的分类模型参数。

15 在上述步骤中，初始化训练数据的分类模型参数可以是预先设置上述决策树的个数、每个决策树的层数，即对决策树的属性进行初步设置。

步骤 S233，提取所述训练数据中的多个特征数据，将每个特征数据作为根节点在创建所述多个决策树，并将每个特征数据对应的特征值作为对应的决策树的叶子节点。

20 步骤 S235，计算当前所有叶子节点的最优划分以及其增益，并以增益最大的叶子节点以及对应的划分点进行分裂，使得将所述样本磁盘数据划分到子节点中。

25 在上述步骤中，增益可以是标签值的最小化均方差，即每个样本的标签值与预测标签值做差后，求的差的平方，并计算所有差的平方的和，可以认为被预测出错的样本越多，均方差就越大，因此通过最小化均方差能够找到最优的分枝依据

上述决策树可以是以每个特征数据作为根节点的二叉树，且每个特数据对应于一个特征值，该特征值为以该特征数据为根节点的决策树的叶子

节点。在确定决策树的叶子节点后，对叶子节点在进行下一步划分，此处值得注意的是，当对叶子节点进行进一步划分时，在多个叶子节点的增益不相同的情况下，划分增益最大的叶子节点，使所有样本数据都能划分至相应的叶子节点中。

5 在一种可选的实施例中，以样本磁盘为 A、B、C 和 D 四块磁盘为例，其中，A 磁盘和 B 磁盘为正常磁盘，C 磁盘和 D 磁盘为损坏的磁盘，在这一示例中，将正常磁盘对应于 0，故障磁盘对应于 1，因此，A、B、C 和 D 四块磁盘分别对应为 0、0、1、1。获取上述磁盘在第一维度上的特征值为 A，使用 GBDT 算法对样本磁盘数据进行训练，图 3 是根据本发明实施
10 例的一种使用 GBDT 算法对样本磁盘数据进行训练的示意图，结合图 3 所示，设置默认初始值为 0.5，即每个磁盘为故障磁盘的概率为 0.5，第一维度的阈值为 A_0 ，将特征值大于 A_0 的磁盘划分为一个子节点，将第一维度上的特征值小于等于 A_0 的磁盘划分为另一个子节点，并设置两个子节点的磁盘为故障磁盘的概率为 0.5。

15 此处需要说明的当是，上述实施例为方便说明，仅选用了四个样本数据进行说明，因此只划分得到两个叶子节点，在实际应用中，根节点划分为两个叶子节点之后，仍可以继续划分，样本数据量越大，划分的层次就越多。

20 根据本申请上述实施例，在一种优选的方案中，提取所述训练数据中的多个特征数据，将每个特征数据作为根节点在创建所述多个决策树，并将每个特征数据对应的特征值作为对应的决策树的叶子节点，包括：

步骤 S2331，读取任意一个特征数据对应的阈值。

步骤 S2333，将所述任意一个特征数据的特征值与所述阈值进行比较，并根据比较结果得到两个分支的熵。

25 步骤 S2335，根据所述两个分支的熵确定两个新节点作为所述任意一个特征数据的两个叶子节点。

步骤 S2337，采用上述步骤对每一个特征数据进行处理，直到每个特

征数据得到预定的两个唯一的叶子节点。

在上述步骤中，穷举每一个特征的每一个阈值，找到使得按照特征小于等于阈值，和特征大于阈值分成的两个分枝的熵最小的特征和阈值，按照该标准分枝得到两个新节点，使用同样方法继续分枝直到所有样本都被分入只有正常磁盘或只有故障磁盘的叶子节点，或达到预设的终止条件，若最终叶子节点中不是只有正常磁盘或故障磁盘，则以该节点上所有样本的平均标签值作为该叶子节点的预测标签值。

此处需要说明的是，标签值即为该磁盘为故障磁盘的概率。

此处仍需要说明的是，熵最小是指尽可能的使每个分枝中，正样本和负样本的比例远离 1:1，熵最小的情况为该分枝上只有正样本或负样本，即该分支上只有正常的磁盘，或故障磁盘。

在一种可选的实施例中，在决策树为回归树的示例中，每个节点都会得一个预测值，该预测值等于属于该节点的所有标签值的平均值，对该节点进行划分时，穷举每一个特征的每个阈值，找最好的分割点进行划分，直到每个叶子节点上每个样本的标签值都唯一或者达到预设的终止条件，若最终叶子节点上样本的标签值不唯一，则以该节点上所有样本的平均标签值作为该叶子节点的预测标签值。

此处需要说明的是，在上述实施例中，最优的划分标准不再是最小化熵，而是最小化均方差，即每个样本的标签值与预测标签值做差后，求的差的平方，并计算所有差的平方的和，可以认为被预测出错的样本越多，均方差就越大，因此通过最小化均方差能够找到最优的分枝依据。

此处还需要说明的是，在进行划分时，使每个叶子节点上每个样本的标签值都唯一是很难达到的，因此为了得到最接近真实情况的预测结果可以预设一个终止条件，该终止条件可以是叶子的上限。

根据本申请上述实施例，在一种优选的方案中，在得到由多个决策树组成的磁盘预测模型之后，所述方法还包括：对所述分类模型参数进行调整，其中，在所述分类模型参数包括故障磁盘样本和非故障磁盘样本的情

况下，如果要确定所述待测磁盘是否为故障磁盘，则将所述分类模型参数中的故障磁盘样本的比例调高。

根据本申请上述实施例，在一种优选的方案中，使用所述由多个决策树组成的磁盘预测模型对所述待测磁盘的磁盘数据进行处理，确定所述待测磁盘是否为故障磁盘，包括：

步骤 S251，接收到所述待测磁盘的磁盘数据之后，对所述待测磁盘的磁盘数据赋予一个初始值。

步骤 S253，根据所述待测磁盘的初始值遍历每一个决策树，计算得到第一个决策树所确定的预测结果和第一残差，并将所述第一残差赋值给所述初始值，得到更新后的初始值。

步骤 S255，以所述更新后的初始值计算得到第二个决策树所确定的预测结果和第二残差，并将所述第二残差赋值所述更新后的初始值，以此遍历所有的决策树，得到预测所述待测磁盘是否为故障磁盘的结果。

步骤 S257，每一棵树学的是之前所有树结论和的残差，这个残差就是一个加预测值后能得真实值的累加量。

在一种可选的实施例中，仍以上述 A,B,C,D 四个磁盘为例，采用特征 A 可将 A,B,C,D 四个磁盘分为两个部分，分别为 A,B 和 C,D，每个部分用平均标签值作为预测值。此时计算残差，其中残差至为磁盘的预测值与磁盘的实际值的差，所以 A 的残差就是 $1-0.5=0.5$ 进而得到 A,B,C,D 的残差分别为 0.5,-0.5, 0.5,-0.5。然后结合图 4 所示，图 4 为根据本发明实施例的一种使用 GBDT 算法计算磁盘预测值的示意图，使用残差替代 A,B,C,D 的原值，输入至第二棵决策树进行训练，并根据与特征 B 的比对结果分为两个叶子节点，如果预测值和它们的残差相等，则只需把第二棵树的结论累加到第一棵树上就能得到磁盘的实际值。第二棵树仅有两个值 0.5 和 -0.5，因此直接分成两个节点。此时所有人的残差都是 0，即每个人都得到了真实的预测值。

此处需要说明的是，上述实施例以说明为目的，因此只有两颗决策树，

在实际应用中，根据样本数据量可以获得多个决策树，且预测值是指之前所有树累加的和，由于此实施例中，这棵决策树之前仅有一颗决策树，因此直接是 0.5，如果还有奇特决策树，则需要都累加起来作为 A 的预测值。

图 5 是根据本发明实施例的一种可选的磁盘的故障预测方法的流程图，
5 下面结合图 5 详细介绍本申请的一种优选的实施例。

如图 5 所示，提供了一种磁盘的故障预测方法，该方法可以包括如下步骤 S51 至步骤 S57：

S51，获取样本磁盘的样本数据。

具体的，在上述步骤中，可以通过 HDTune、CrystalDiskInfo 等软件获
10 取样本磁盘数据。

S52，对样本数据进行差分运算。

具体的，在上述步骤中，差分运算指磁盘在某一时刻的特征数据与过
该磁盘在 24 小时之前的特征数据做差运算得到的值。

S53，对差分运算得到的结果进行分布求和和/或平方运算。

15 S54，得到训练和预测数据。

S55，第一步训练和预测，使召回率较大。

S56，第二步训练和预测，平衡召回率和准确率。

具体的，在上述步骤中，由于训练数据中负样本占比很大，正样本占
比小，例如，当二者比例为 1000:1 时，如果用全部的训练数据做训练，能
20 准确预测的正样本是很少的，由于训练数据中正样本较少，很多真实值为
负样本的数据可能被误判为正样本，因此第一步在训练时使正样本的召回
率较大，第二步在训练时，把第一步预测为正样本的训练数据作为第二步
的训练数据，即选择为与正样本接近的那些样本作为训练样本，如此在做
训练时，训练出的模型会更有利于预测出正样本，这样第二步预测得到的
25 结果，正样本的准确率会比第一步有大幅度提高，从而使准确率和召回率
达到一定的平衡程度。

实施例 2

根据本发明实施例，还提供了一种用于实施上述反编译数据的处理方法的反编译数据的处理装置，图 6 是根据本发明实施例的一种磁盘的故障预测装置的结构示意图，如图 6 所示，该装置包括：获取模块 60，训练模块 62 和处理模块 64。

获取模块 60，用于通过磁盘监控技术获取磁盘的样本磁盘数据，其中，所述样本磁盘数据包括多个维度上的样本数据；

训练模块 62，用于采用 GBDT 算法对所述样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型；

处理模块 64，在接收到待测磁盘的磁盘数据之后，使用所述由多个决策树组成的磁盘预测模型对所述待测磁盘的磁盘数据进行处理，确定所述待测磁盘是否为故障磁盘。

此处需要说明的是，上述获取模块 60，训练模块 62 和处理模块 64 对应于实施例一种的步骤 S21 至步骤 S25 所实现的实例和应用场景相同，但不限于上述实施例一所公开的内容。需要说明的是，上述模块作为装置的一部分可以运行在实施例一提供的计算机终端 10 中。

根据本申请上述实施例，在一种优选的方案中，所述样本磁盘数据为 SMART 磁盘数据，其中，所述样本磁盘数据至少包括如下四个维度上的样本数据：原始值、标准值、最差值和累积值。

根据本申请上述实施例，在一种优选的方案中，结合图 7 所示，上述装置还包括：

运算模块 70，用于对所述每个维度上的样本数据进行如下任意一种或多种运算：差分运算、平方运算和分布求和运算，使得任意一个维度上的样本数据被扩展出新的维度上的样本数据。

此处需要说明的是，上述运算模块 770 对应与实施例一中的步骤 S21 至步骤 S25 所实现的实例和应用场景相同，但不限于上述实施例一所公开的内容。需要说明的是，上述模块作为装置的一部分可以运行在实施例一

提供的计算机终端 10 中。

根据本申请上述实施例，在一种优选的方案中，结合图 8 所示，上述训练模块 62 还包括：

5 初始模块 80，用于以所有磁盘的样本磁盘数据作为训练数据，并采用默认值初始化所述训练数据的分类模型参数；

提取模块 82，用于提取所述训练数据中的多个特征数据，将每个特征数据作为根节点在创建所述多个决策树，并将每个特征数据对应的特征值作为对应的决策树的叶子节点；

10 第一计算模块 84，用于计算当前所有叶子节点的最优划分以及其增益，并以增益最大的叶子节点以及对应的划分点进行分裂，使得将所述样本磁盘数据划分到子节点中。

15 此处需要说明的是，上述初始模块 80，提取模块 82 和第一计算模块 84 对应于实施例一种的步骤 S231 至步骤 S235 所实现的实例和应用场景相同，但不限于上述实施例一所公开的内容。需要说明的是，上述模块作为装置的一部分可以运行在实施例一提供的计算机终端 10 中。

根据本申请上述实施例，在一种优选的方案中，结合图 9 所示，所述提取模块 82 包括：

读取模块 90，用于读取任意一个特征数据对应的阈值；

20 比较模块 92，用于将所述任意一个特征数据的特征值与所述阈值进行比较，并根据比较结果得到两个分支的熵；

确定模块 94，用于根据所述两个分支的熵确定两个新节点作为所述任意一个特征数据的两个叶子节点；

处理子模块 96，用于采用上述步骤对每一个特征数据进行处理，直到每个特征数据得到预定的两个唯一的叶子节点。

25 此处需要说明的是，上述读取模块 90，比较模块 92、确定模块 94 和处理子模块 96 对应于实施例一种的步骤 S2331 至步骤 S2337 所实现的实

例和应用场景相同，但不限于上述实施例一所公开的内容。需要说明的是，上述模块作为装置的一部分可以运行在实施例一提供的计算机终端 10 中。

根据本申请上述实施例，在一种优选的方案中，在得到由多个决策树组成的磁盘预测模型之后，所述方法还包括：对所述分类模型参数进行调整，其中，在所述分类模型参数包括故障磁盘样本和非故障磁盘样本的情况下，如果要确定所述待测磁盘是否为故障磁盘，则将所述分类模型参数中的故障磁盘样本的比例调高。

根据本申请上述实施例，在一种优选的方案中，结合图 10 所示，上述处理模块 64 包括：

10 接收模块 100，用于接收到所述待测磁盘的磁盘数据之后，对所述待测磁盘的磁盘数据赋予一个初始值；

第二计算模块 102，用于根据所述待测磁盘的初始值遍历每一个决策树，计算得到第一个决策树所确定的预测结果和第一残差，并将所述第一残差赋值给所述初始值，得到更新后的初始值；

15 遍历模块 104，用于以所述更新后的初始值计算得到第二个决策树所确定的预测结果和第二残差，并将所述第二残差赋值所述更新后的初始值，以此遍历所有的决策树，得到预测所述待测磁盘是否为故障磁盘的结果。

此处需要说明的是，上述接收模块 100，第二计算模块 102 和遍历模块 104 对应于实施例一种的步骤 S251 至步骤 S255 所实现的实例和应用场景相同，但不限于上述实施例一所公开的内容。需要说明的是，上述模块作为装置的一部分可以运行在实施例一提供的计算机终端 10 中。

实施例 3

本发明的实施例可以提供一种计算机终端，该计算机终端可以是计算机终端群中的任意一个计算机终端设备。可选地，在本实施例中，上述计
25 算机终端也可以替换为移动终端等终端设备。

可选地，在本实施例中，上述计算机终端可以位于计算机网络的多个网络设备中的至少一个网络设备。

在本实施例中，上述计算机终端可以执行磁盘的故障预测方法中以下步骤的程序代码：通过磁盘监控技术获取磁盘的样本磁盘数据，其中，样本磁盘数据包括多个维度上的样本数据；采用 GBDT 算法对样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型；在接收到待测磁
5 盘的磁盘数据之后，使用由多个决策树组成的磁盘预测模型对待测磁盘的磁盘数据进行处理，确定待测磁盘是否为故障磁盘。

可选地，图 11 是根据本发明实施例的一种计算机终端的结构框图。如图 11 所示，该计算机终端 A 可以包括：一个或多个（图中仅示出一个）处理器 111、存储器 113、以及传输装置 115。

10 其中，存储器可用于存储软件程序以及模块，如本发明实施例中的磁盘的故障预测方法和装置对应的程序指令/模块，处理器通过运行存储在存储器内的软件程序以及模块，从而执行各种功能应用以及数据处理，即实现上述的磁盘的故障预测方法。存储器可包括高速随机存储器，还可以包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易
15 失性固态存储器。在一些实例中，存储器可进一步包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至终端 A。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

处理器可以通过传输装置调用存储器存储的信息及应用程序，以执行下述步骤：样本磁盘数据为 SMART 磁盘数据，其中，样本磁盘数据至少
20 包括如下四个维度上的样本数据：原始值、标准值、最差值和累积值。

可选的，上述处理器还可以执行如下步骤的程序代码：对每个维度上的样本数据进行如下任意一种或多种运算：差分运算、平方运算和分布求和运算，使得任意一个维度上的样本数据被扩展出新的维度上的样本数据。

可选的，上述处理器还可以执行如下步骤的程序代码：以所有磁盘的
25 样本磁盘数据作为训练数据，并采用默认值初始化训练数据的分类模型参数；提取训练数据中的多个特征数据，将每个特征数据作为根节点在创建多个决策树，并将每个特征数据对应的特征值作为对应的决策树的叶子节点；计算当前所有叶子节点的最优划分以及其增益，并以增益最大的叶子

节点以及对应的划分点进行分裂，使得将样本磁盘数据划分到子节点中。

可选的，上述处理器还可以执行如下步骤的程序代码：读取任意一个特征数据对应的阈值；将任意一个特征数据的特征值与阈值进行比较，并根据比较结果得到两个分支的熵；根据两个分支的熵确定两个新节点作为任意一个特征数据的两个叶子节点；采用上述步骤对每一个特征数据进行处理，直到每个特征数据得到预定的两个唯一的叶子节点。

可选的，上述处理器还可以执行如下步骤的程序代码：在得到由多个决策树组成的磁盘预测模型之后，方法还包括：对分类模型参数进行调整，其中，在分类模型参数包括故障磁盘样本和非故障磁盘样本的情况下，如果要确定待测磁盘是否为故障磁盘，则将分类模型参数中的故障磁盘样本的比例调高。

可选的，上述处理器还可以执行如下步骤的程序代码：接收到待测磁盘的磁盘数据之后，对待测磁盘的磁盘数据赋予一个初始值；根据待测磁盘的初始值遍历每一个决策树，计算得到第一个决策树所确定的预测结果和第一残差，并将第一残差赋值给初始值，得到更新后的初始值；以更新后的初始值计算得到第二个决策树所确定的预测结果和第二残差，并将第二残差赋值更新后的初始值，以此遍历所有的决策树，得到预测待测磁盘是否为故障磁盘的结果。

在本发明实施例中，采用通过磁盘监控技术获取磁盘的样本磁盘数据，其中，样本磁盘数据包括多个维度上的样本数据；采用 GBDT 算法对样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型方式，通过在接收到待测磁盘的磁盘数据之后，使用由多个决策树组成的磁盘预测模型对待测磁盘的磁盘数据进行处理，达到了确定待测磁盘是否为故障磁盘的目的，从而实现了预测磁盘故障状态的技术效果，进而解决了现有技术的硬盘故障预测系统中一些容易致使硬盘故障的因素不能被采集或量化导致的预测结果不准确的技术问题。

本领域普通技术人员可以理解，图 11 所示的结构仅为示意，计算机终端也可以是智能手机（如 Android 手机、iOS 手机等）、平板电脑、掌上电

脑以及移动互联网设备(Mobile Internet Devices, MID)、PAD 等终端设备。

图 11 其并不对上述电子装置的结构造成限定。例如, 计算机终端 A 还可包括比图 11 中所示更多或者更少的组件(如网络接口、显示装置等), 或者具有与图 11 所示不同的配置。

- 5 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通程序来指令终端设备相关的硬件来完成, 该程序可以存储于一计算机可读存储介质中, 存储介质可以包括: 闪存盘、只读存储器(Read-Only Memory, ROM)、随机存取器(Random Access Memory, RAM)、磁盘或光盘等。

10 **实施例 4**

本发明的实施例还提供了一种存储介质。可选地, 在本实施例中, 上述存储介质可以用于保存上述实施例一所提供的磁盘的故障预测方法所执行的程序代码。

- 15 可选地, 在本实施例中, 上述存储介质可以位于计算机网络中计算机终端群中的任意一个计算机终端中, 或者位于移动终端群中的任意一个移动终端中。

- 20 可选地, 在本实施例中, 存储介质被设置为存储用于执行以下步骤的程序代码: 通过磁盘监控技术获取磁盘的样本磁盘数据, 其中, 样本磁盘数据包括多个维度上的样本数据; 采用 GBDT 算法对样本磁盘数据进行样本训练, 得到由多个决策树组成的磁盘预测模型; 在接收到待测磁盘的磁盘数据之后, 使用由多个决策树组成的磁盘预测模型对待测磁盘的磁盘数据进行处理, 确定待测磁盘是否为故障磁盘。

- 25 可选地, 上述存储介质还被设置为存储用于执行以下步骤的程序代码: 对每个维度上的样本数据进行如下任意一种或多种运算: 差分运算、平方运算和分布求和运算, 使得任意一个维度上的样本数据被扩展出新的维度上的样本数据。

可选地, 上述存储介质还被设置为存储用于执行以下步骤的程序代码:

以所有磁盘的样本磁盘数据作为训练数据，并采用默认值初始化训练数据的分类模型参数；提取训练数据中的多个特征数据，将每个特征数据作为根节点在创建多个决策树，并将每个特征数据对应的特征值作为对应的决策树的叶子节点；计算当前所有叶子节点的最优划分以及其增益，并以增益最大的叶子节点以及对应的划分点进行分裂，使得将样本磁盘数据划分到子节点中。

可选地，上述存储介质还被设置为存储用于执行以下步骤的程序代码：读取任意一个特征数据对应的阈值；将任意一个特征数据的特征值与阈值进行比较，并根据比较结果得到两个分支的熵；根据两个分支的熵确定两个新节点作为任意一个特征数据的两个叶子节点；采用上述步骤对每一个特征数据进行处理，直到每个特征数据得到预定的两个唯一的叶子节点。

可选地，上述存储介质还被设置为存储用于执行以下步骤的程序代码：在得到由多个决策树组成的磁盘预测模型之后，方法还包括：对分类模型参数进行调整，其中，在分类模型参数包括故障磁盘样本和非故障磁盘样本的情况下，如果要确定待测磁盘是否为故障磁盘，则将分类模型参数中的故障磁盘样本的比例调高。

可选地，上述存储介质还被设置为存储用于执行以下步骤的程序代码：接收到待测磁盘的磁盘数据之后，对待测磁盘的磁盘数据赋予一个初始值；根据待测磁盘的初始值遍历每一个决策树，计算得到第一个决策树所确定的预测结果和第一残差，并将第一残差赋值给初始值，得到更新后的初始值；以更新后的初始值计算得到第二个决策树所确定的预测结果和第二残差，并将第二残差赋值更新后的初始值，以此遍历所有的决策树，得到预测待测磁盘是否为故障磁盘的结果。

上述本发明实施例序号仅仅为了描述，不代表实施例的优劣。

在本发明的上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详述的部分，可以参见其他实施例的相关描述。

在本申请所提供的几个实施例中，应该理解到，所揭露的技术内容，

可通过其它的方式实现。其中，以上所描述的装置实施例仅仅是示意性的，例如所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，单元或模块的间接耦合或通信连接，可以是电性或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本发明各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读取存储介质中。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备（可为个人计算机、服务器或者网络设备）执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

以上所述仅是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

权 利 要 求 书

1. 一种磁盘的故障预测方法，其特征在于，包括：

通过磁盘监控技术获取磁盘的样本磁盘数据，其中，所述样本磁盘数据包括多个维度上的样本数据；

采用 GBDT 算法对所述样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型；

在接收到待测磁盘的磁盘数据之后，使用所述由多个决策树组成的磁盘预测模型对所述待测磁盘的磁盘数据进行处理，确定所述待测磁盘是否为故障磁盘。

2. 根据权利要求 1 所述的方法，其特征在于，所述样本磁盘数据为 SMART 磁盘数据，其中，所述样本磁盘数据至少包括如下四个维度上的样本数据：原始值、标准值、最差值和累积值。

3. 根据权利要求 2 所述的方法，其特征在于，在通过磁盘监控技术获取磁盘的样本磁盘数据之后，所述方法还包括：

对每个维度上的样本数据进行如下任意一种或多种运算：差分运算、平方运算和分布求和运算，使得任意一个维度上的样本数据被扩展出新的维度上的样本数据。

4. 根据权利要求 1 至 3 中任意一项所述的方法，其特征在于，采用 GBDT 算法对所述样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型，包括：

以所有磁盘的样本磁盘数据作为训练数据，并采用默认值初始化所述训练数据的分类模型参数；

提取所述训练数据中的多个特征数据，将每个特征数据作为根节点在创建所述多个决策树，并将每个特征数据对应的特征值作为对应的决策树的叶子节点；

计算当前所有叶子节点的最优划分以及其增益，并以增益最大的叶子节点以及对应的划分点进行分裂，使得将所述样本磁盘数据划分到子节点中。

5. 根据权利要求 4 所述的方法，其特征在于，提取所述训练数据中的多个特征数据，将每个特征数据作为根节点在创建所述多个决策树，并将每个特征数据对应的特征值作为对应的决策树的叶子节点，包括：

读取任意一个特征数据对应的阈值；

将所述任意一个特征数据的特征值与所述阈值进行比较，并根据比较结果得到两个分支的熵；

根据所述两个分支的熵确定两个新节点作为所述任意一个特征数据的两个叶子节点；

采用上述步骤对每一个特征数据进行处理，直到每个特征数据得到预定的两个唯一的叶子节点。

6. 根据权利要求 4 所述的方法，其特征在于，在得到由多个决策树组成的磁盘预测模型之后，所述方法还包括：对所述分类模型参数进行调整，其中，在所述分类模型参数包括故障磁盘样本和非故障磁盘样本的情况下，如果要确定所述待测磁盘是否为故障磁盘，则将所述分类模型参数中的故障磁盘样本的比例调高。

7. 根据权利要求 1 所述的方法，其特征在于，使用所述由多个决策树组成的磁盘预测模型对所述待测磁盘的磁盘数据进行处理，确定所述待测磁盘是否为故障磁盘，包括：

接收到所述待测磁盘的磁盘数据之后，对所述待测磁盘的磁盘数据赋予一个初始值；

根据所述待测磁盘的初始值遍历每一个决策树，计算得到第一个决策树所确定的预测结果和第一残差，并将所述第一残差赋值给所述初始值，得到更新后的初始值；

以所述更新后的初始值计算得到第二个决策树所确定的预测结果和第二残差，并所述第二残差赋值所述更新后的初始值，以此遍历所有的决策树，得到预测所述待测磁盘是否为故障磁盘的结果。

8. 一种磁盘的故障预测装置，其特征在于，包括：

5 获取模块，用于通过磁盘监控技术获取磁盘的样本磁盘数据，其中，所述样本磁盘数据包括多个维度上的样本数据；

 训练模块，用于采用 GBDT 算法对所述样本磁盘数据进行样本训练，得到由多个决策树组成的磁盘预测模型；

10 处理模块，在接收到待测磁盘的磁盘数据之后，使用所述由多个决策树组成的磁盘预测模型对所述待测磁盘的磁盘数据进行处理，确定所述待测磁盘是否为故障磁盘。

9. 根据权利要求 8 所述的装置，其特征在于，所述样本磁盘数据为 SMART 磁盘数据，其中，所述样本磁盘数据至少包括如下四个维度上的样本数据：原始值、标准值、最差值和累积值。

15 10. 根据权利要求 9 所述的装置，其特征在于，所述装置还包括：

 运算模块，用于对每个维度上的样本数据进行如下任意一种或多种运算：差分运算、平方运算和分布求和运算，使得任意一个维度上的样本数据被扩展出新的维度上的样本数据。

20 11. 根据权利要求 8 至 10 中任意一项所述的装置，其特征在于，所述训练模块还包括：

 初始模块，用于以所有磁盘的样本磁盘数据作为训练数据，并采用默认值初始化所述训练数据的分类模型参数；

25 提取模块，用于提取所述训练数据中的多个特征数据，将每个特征数据作为根节点在创建所述多个决策树，并将每个特征数据对应的特征值作为对应的决策树的叶子节点；

 第一计算模块，用于计算当前所有叶子节点的最优划分以及其增

益，并以增益最大的叶子节点以及对应的划分点进行分裂，使得将所述样本磁盘数据划分到子节点中。

12. 根据权利要求 11 所述的装置，其特征在于，所述提取模块包括：

读取模块，用于读取任意一个特征数据对应的阈值；

5 比较模块，用于将所述任意一个特征数据的特征值与所述阈值进行比较，并根据比较结果得到两个分支的熵；

确定模块，用于根据所述两个分支的熵确定两个新节点作为所述任意一个特征数据的两个叶子节点；

10 处理子模块，用于采用上述步骤对每一个特征数据进行处理，直到每个特征数据得到预定的两个唯一的叶子节点。

13. 根据权利要求 11 所述的装置，其特征在于，在得到由多个决策树组成的磁盘预测模型之后，所述装置还包括：对所述分类模型参数进行调整，其中，在所述分类模型参数包括故障磁盘样本和非故障磁盘样本的情况下，如果要确定所述待测磁盘是否为故障磁盘，则将所述分类模型参数中的故障磁盘样本的比例调高。

15

14. 根据权利要求 8 所述的装置，其特征在于，所述处理模块包括：

接收模块，用于接收到所述待测磁盘的磁盘数据之后，对所述待测磁盘的磁盘数据赋予一个初始值；

20 第二计算模块，用于根据所述待测磁盘的初始值遍历每一个决策树，计算得到第一个决策树所确定的预测结果和第一残差，并将所述第一残差赋值给所述初始值，得到更新后的初始值；

25 遍历模块，用于以所述更新后的初始值计算得到第二个决策树所确定的预测结果和第二残差，并将所述第二残差赋值所述更新后的初始值，以此遍历所有的决策树，得到预测所述待测磁盘是否为故障磁盘的结果。

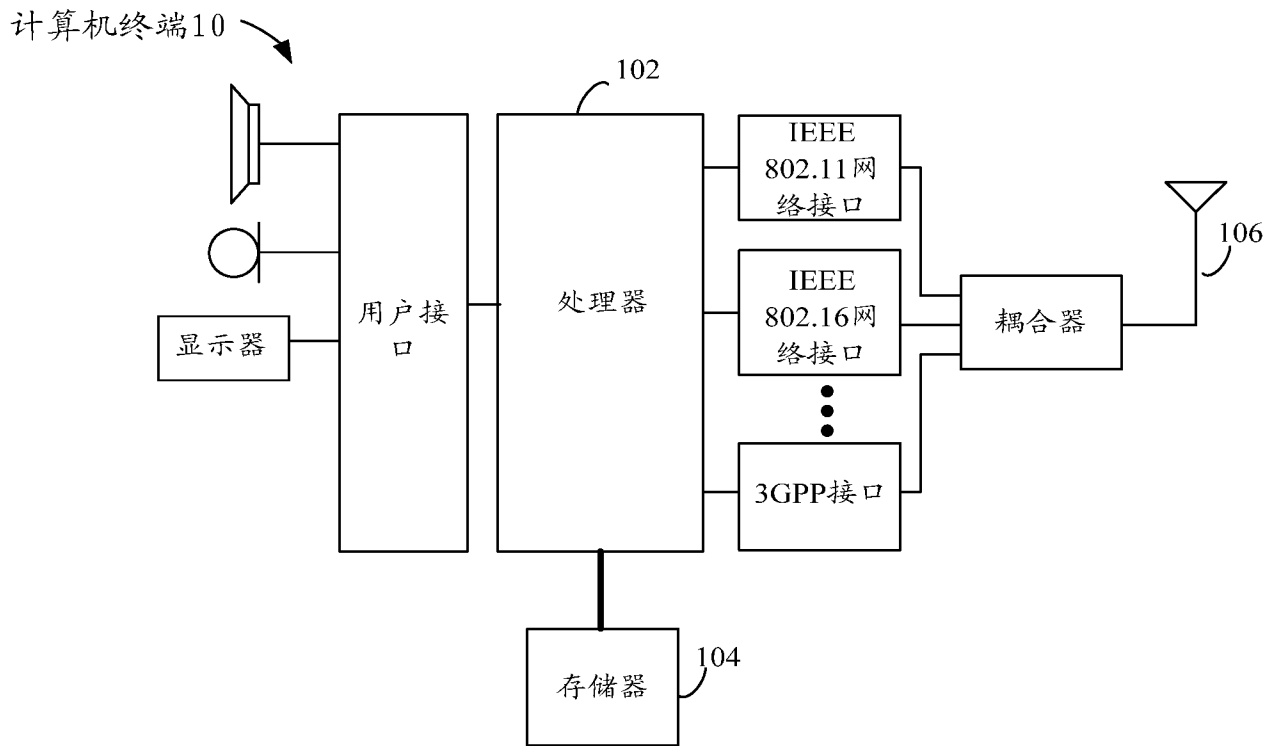


图 1

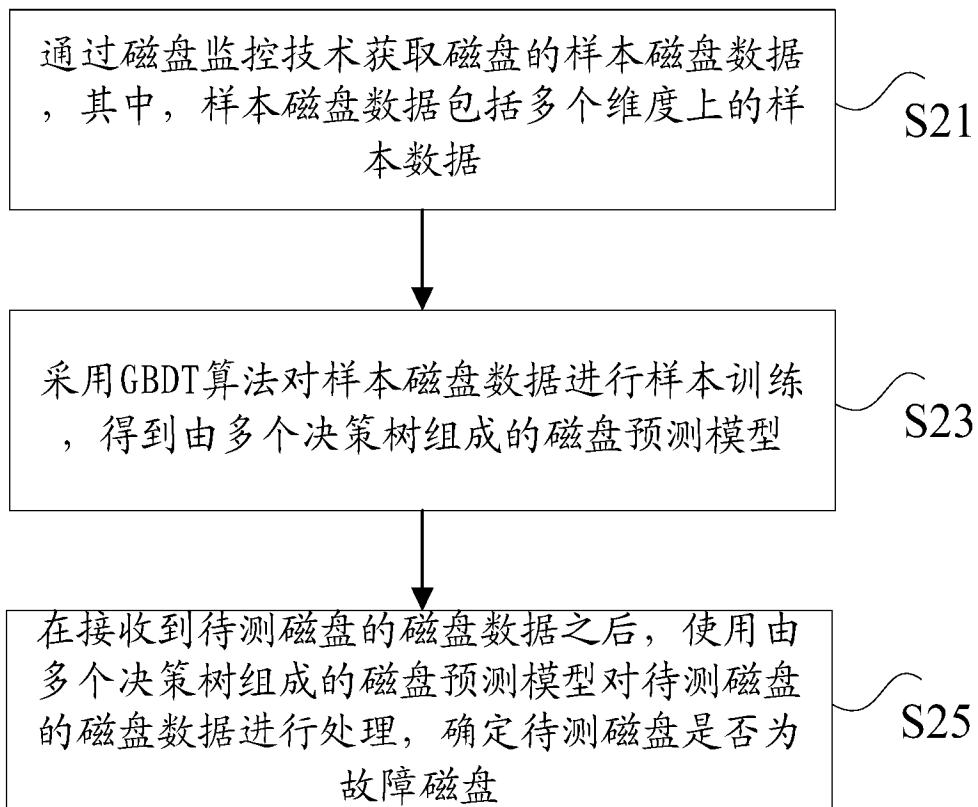


图 2

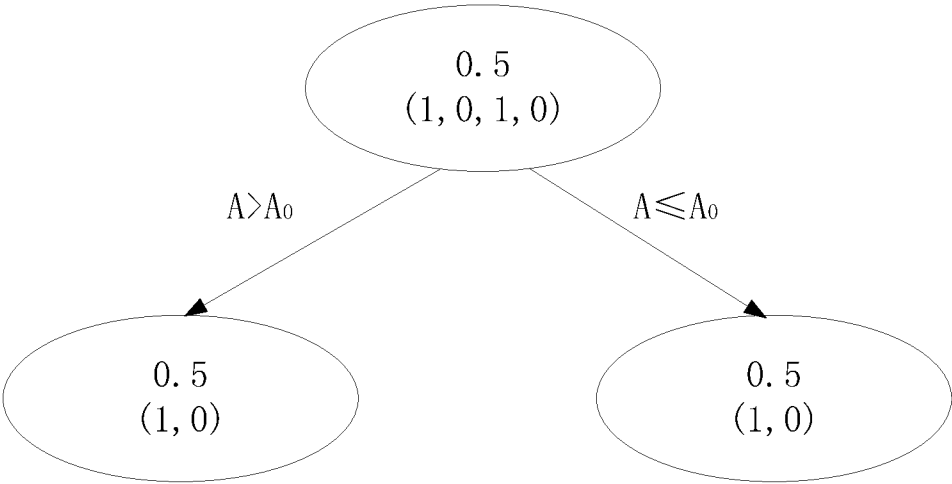


图 3

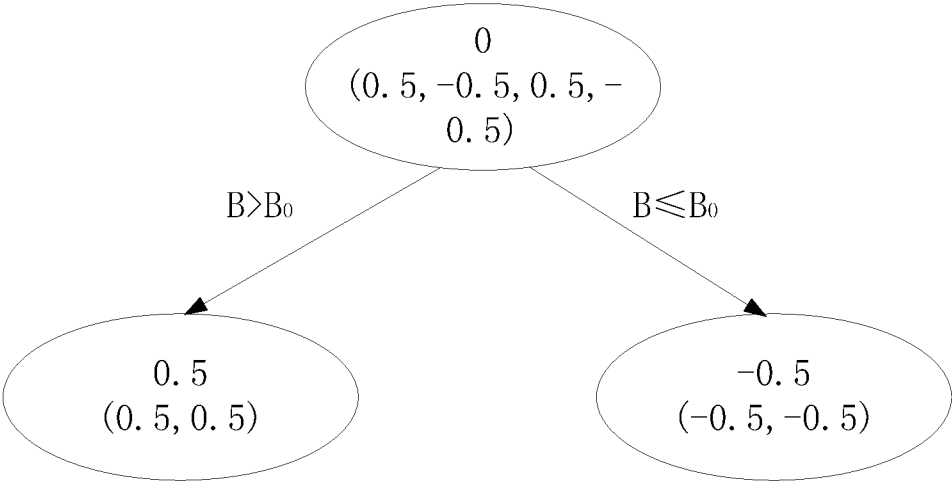


图 4

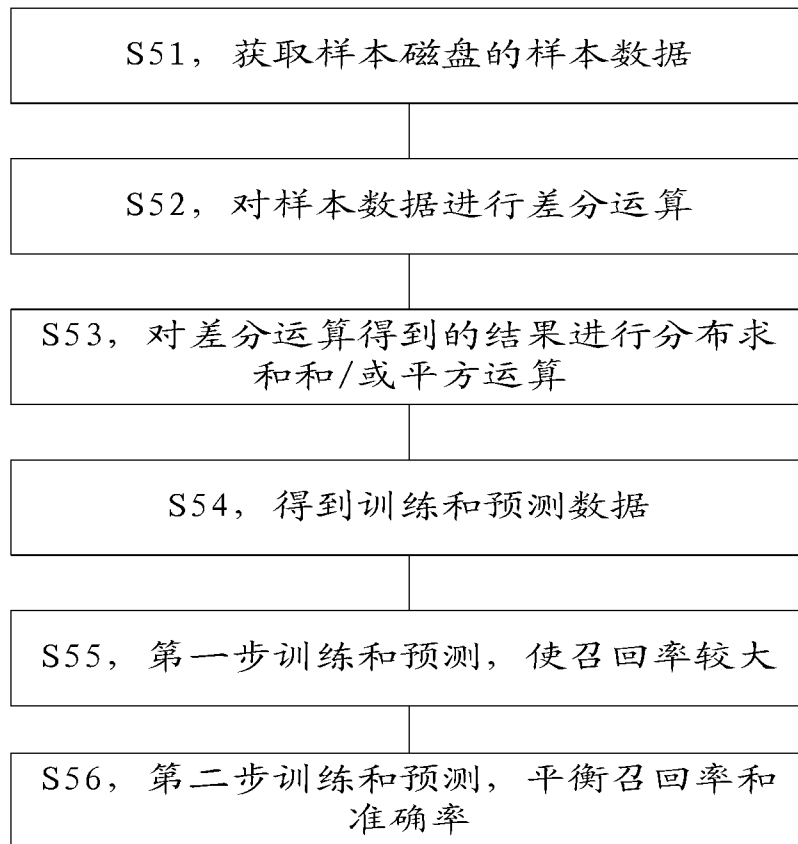


图 5



图 6



图 7

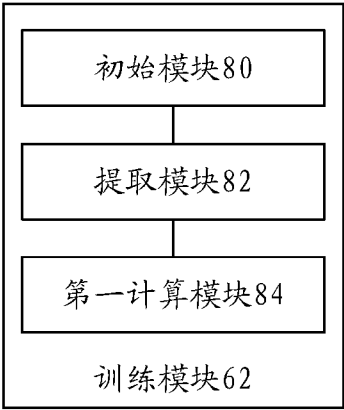


图 8

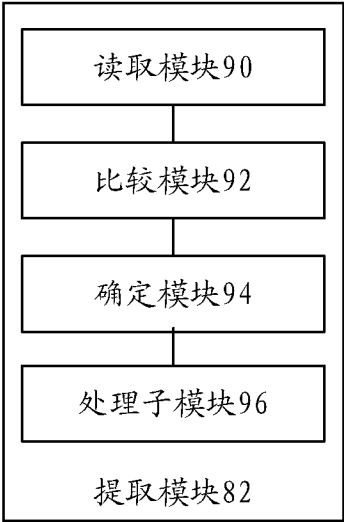


图 9

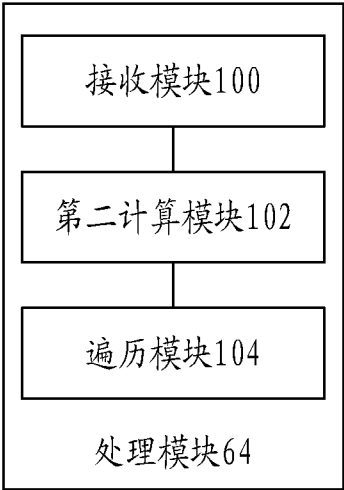


图 10

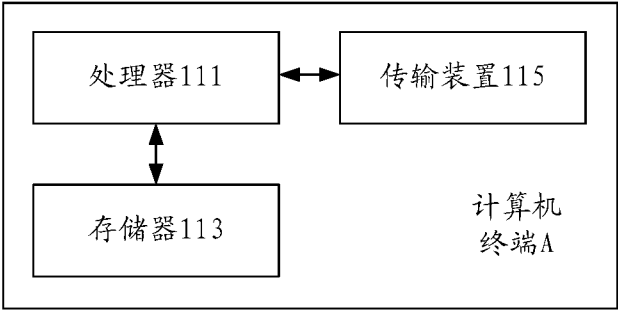


图 11

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/071695

A. CLASSIFICATION OF SUBJECT MATTER

G06F 11/22 (2006.01) 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

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)

WPI, EPODOC, CNPAT, CNKI: hard disc, decision tree, promote, damage, disk, hard, medium, storage, decision, tree, gradient, boost, tree-link, sample, model, training, fault, error, problem, prediction, monitor, detect

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 105224888 A (EISOO SOFTWARE CO., LTD.), 06 January 2016 (06.01.2016), abstract, description, paragraphs [0046]-[0058], and figure 3	1-14
Y	CN 104766014 A (ANYIHENG TONG (BEIJING) TECHNOLOGY CO., LTD.), 08 July 2015 (08.07.2015), description, paragraphs [0087]-[0088]	1-14
A	CN 101866271 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY), 20 October 2010 (20.10.2010), the whole document	1-14
A	CN 105069476 A (STATE GRID NINGXIA ELECTRIC POWER CO., LTD. et al.), 18 November 2015 (18.11.2015), the whole document	1-14
A	US 2015205657 A1 (LONGSAND LIMITED), 23 July 2015 (23.07.2015), the whole document	1-14

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 14 April 2016 (14.04.2016)	Date of mailing of the international search report 27 April 2017 (27.04.2017)
Name and mailing address of the ISA/CN: 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 DAI, Lei Telephone No.: (86-10) 62414073

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2017/071695

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105224888 A	06 January 2016	None	
CN 104766014 A	08 July 2015	WO 2016173200 A1	03 November 2016
CN 101866271 A	20 October 2010	None	
CN 105069476 A	18 November 2015	None	
US 2015205657 A1	23 July 2015	EP 2901284 A1	05 August 2015
		WO 2014051603 A1	03 April 2014

国际检索报告

国际申请号

PCT/CN2017/071695

A. 主题的分类

G06F 11/22 (2006.01) 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

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))

WPI, EPODOC, CNPAT, CNKI; 磁盘, 硬盘, 存储, 介质, 决策树, 梯度, 提高, 提升, 样本, 采样, 模型, 训练, 损坏, 故障, 错误, 问题, 预测, 监测, 检测, disk, hard, medium, storage, decision, tree, gradient, boost, tree-link, sample, model, training, fault, error, problem, prediction, monitor, detect

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 105224888 A (上海爱数软件有限公司) 2016年 1月 6日 (2016 - 01 - 06) 摘要, 说明书第[0046]-[0058]段, 附图3	1-14
Y	CN 104766014 A (安一恒通北京科技有限公司) 2015年 7月 8日 (2015 - 07 - 08) 说明书第[0087]-[0088]段	1-14
A	CN 101866271 A (华中科技大学) 2010年 10月 20日 (2010 - 10 - 20) 全文	1-14
A	CN 105069476 A (国网宁夏电力公司 等) 2015年 11月 18日 (2015 - 11 - 18) 全文	1-14
A	US 2015205657 A1 (LONGSAND LIMITED) 2015年 7月 23日 (2015 - 07 - 23) 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

2017年 4月 14日

国际检索报告邮寄日期

2017年 4月 27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

戴雷

电话号码 (86-10) 62414073

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/071695

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	105224888	A	2016年 1月 6日	无			
CN	104766014	A	2015年 7月 8日	WO	2016173200	A1	2016年 11月 3日
CN	101866271	A	2010年 10月 20日	无			
CN	105069476	A	2015年 11月 18日	无			
US	2015205657	A1	2015年 7月 23日	EP	2901284	A1	2015年 8月 5日
				WO	2014051603	A1	2014年 4月 3日