(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103646114 B
(45)授权公告日 2017.04.05

(21)申请号 201310733574.5

(22)申请日 2013.12.26

(65)同一申请的已公布的文献号
申请公布号 CN 103646114 A

(43)申请公布日 2014.03.19

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 胡光 胡殿明 杨文君 魏伟

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 宋合成

(51)Int.Cl.
G06F 17/30(2006.01)

(56)对比文件
US 2013/0293981 A1,2013.11.07,1-14.
CN 1627277 A,2005.06.15,全文.
PCFAN评测室.固态硬盘当缓存 Intel Smart Response技术实战.《电脑迷 》.2011,32-33.

审查员 朱琦

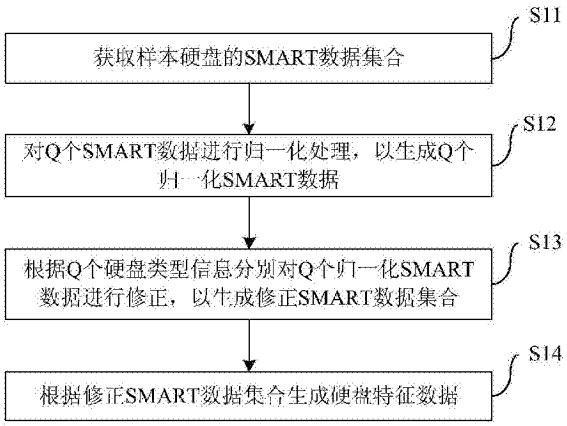
权利要求书4页 说明书9页 附图3页

(54)发明名称

硬盘SMART数据中特征数据提取方法和装置

(57)摘要

本发明提出一种硬盘SMART自我监测、分析和报告技术数据中特征数据提取方法和装置，其中，该方法包括以下步骤：获取样本硬盘的SMART数据集合，其中，SMART数据集合包括Q个SMART数据和与Q个SMART数据分别对应的Q个硬盘类型信息；对Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据；根据Q个硬盘类型信息分别对Q个归一化SMART数据进行修正，以生成修正SMART数据集合；根据修正SMART数据集合生成硬盘特征数据。本发明实施例的方法，可通过同一个故障预警模型实现对不同硬盘的故障预警测试和分析，提高了故障预警模型的准确性，降低了模型训练、测试和分析成本。



CN 103646114 B

1.一种硬盘SMART自我监测、分析和报告技术数据中特征数据提取方法，其特征在于，包括：

获取样本硬盘的SMART数据集合，其中，所述SMART数据集合包括Q个SMART数据和与所述Q个SMART数据分别对应的Q个硬盘类型信息；

对所述Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据；

根据所述Q个硬盘类型信息分别对所述Q个归一化SMART数据进行修正，以生成修正SMART数据集合；

根据所述修正SMART数据集合生成硬盘特征数据；

其中，所述SMART数据集合包括与S个属性分别对应的S个属性数据集合，每个所述SMART数据中包括与S个属性分别对应的S个第一属性数据子集合，在所述对所述Q个SMART数据进行归一化处理之前，还包括：

分别获取每个所述SMART数据中的S个第一属性数据子集合对应的S个梯度数据集合；

将获取的每个所述SMART数据的S个梯度数据集合作为S个新的第一属性数据子集合分别加入所述每个SMART数据。

2.如权利要求1所述的方法，其特征在于，分别获取每个所述SMART数据中的S个第一属性数据子集合对应的S个梯度数据集合具体包括：

从每个所述SMART数据中的第s个属性对应的第一属性数据子集合中依次选取M个属性数据，以生成P个第二属性数据子集合，其中，P＝N-M+1，N为所述属性s对应的第一属性数据子集合中属性数据的总数，s＝1…S；

分别计算所述P个第二属性数据子集合对应的P个梯度数据，并根据所述P个梯度数据生成所述属性s对应的梯度数据集合。

3.如权利要求2所述的方法，其特征在于，所述分别计算所述P个第二属性数据子集合对应的P个梯度数据具体包括：

通过加权最小二乘法分别获取所述P个第二属性数据子集合对应的P个拟合系数，其中，所述P个拟合系数中第i个拟合系数k_i＝(Z-b*Y)/X，

其中，i＝1…P，

$$X=\sum_{j=i}^{M+i-1} w_j*x_j^2,$$

$$Y=\sum_{j=i}^{M+i-1} w_j*x_i,$$

$$Z=\sum_{j=i}^{M+i-1} w_j*x_j*y_j,$$

$$b=(Z*Y-\sum_{j=i}^{M+i-1}(w_j*y_j)*X)/(Y*Y-X*\sum_{j=i}^{M+i-1} w_j),$$

w_j为所述属性s对应的第一属性数据子集合中第j个属性数据对应的预设权重，x_j为所述属性s对应的第一属性数据子集合中第j个属性数据的检测时间，y_j为所述属性s对应的第一属性数据子集合中第j个属性数据；

通过以下公式分别获取所述P个梯度数据中的第i个梯度数据$Grad_i$：

$Grad_i = k_i * (M-1) * y_{M+i-1}$。

4.如权利要求1-3任一项所述的方法，其特征在于，所述对所述Q个SMART数据进行归一化处理具体包括：

通过以下公式对所述Q个SMART数据进行归一化处理：

$g(x) = sign(x) \times \log_y |x|$，

其中，x为所述Q个SMART数据中的一个属性数据，g(x)为属性数据x对应的归一化后的属性数据，所述y可通过以下公式计算：

$y^z \leq Value < (y+\Delta y)^z$，

其中，z为预设阈值，Value为所述属性数据x对应的属性的出厂默认值，Δy为预设精度。

5.如权利要求1-3任一项所述的方法，其特征在于，所述根据所述Q个硬盘类型信息分别对所述Q个归一化SMART数据进行修正具体包括：

根据所述Q个硬盘类型信息获取与所述Q个硬盘类型信息分别对应的Q个修正值；

根据与所述Q个硬盘类型信息分别对应的Q个修正值分别对相应的归一化SMART数据进行修正。

6.如权利要求1-3任一项所述的方法，其特征在于，所述修正SMART数据集合中包括与所述S个属性分别对应的S个修正属性数据集合，所述根据所述修正SMART数据集合生成硬盘特征数据具体包括：

分别获取所述S个属性对应的S个训练特征数据；

分别对每个训练特征数据中的训练特征值进行排序以生成与所述S个属性对应的S个特征序列(V_i)；

通过以下预设映射规则获取每个属性对应的修正属性数据集合中的每个属性值v的特征值f(v)：

$$\begin{cases} f(v)=V_i & V_i \leq v \leq [(V_{i+1}-V_i)/2] \\ f(v)=V_i+1 & V_i+[(V_{i+1}-V_i)/2]+1 < v < V_{i+1} \end{cases}$$；

根据获取的每个属性对应的修正属性数据集合中的每个属性值的特征值生成所述硬盘特征数据。

7.一种硬盘SMART数据中特征数据提取装置，其特征在于，包括：

第一获取模块，用于获取样本硬盘的SMART数据集合，其中，所述SMART数据集合包括Q个SMART数据和与所述Q个SMART数据分别对应的Q个硬盘类型信息；

第一生成模块，用于对所述Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据；

修正模块，用于根据所述Q个硬盘类型信息分别对所述Q个归一化SMART数据进行修正，以生成修正SMART数据集合；

第二生成模块，用于根据所述修正SMART数据集合生成硬盘特征数据；

其中，所述SMART数据集合包括与S个属性分别对应的S个属性数据集合，每个所述SMART数据中包括与S个属性分别对应的S个第一属性数据子集合，在所述第一生成模块之前，还包括：

第二获取模块，用于分别获取每个所述SMART数据中的S个第一属性数据子集合对应的

S个梯度数据集合；

加入模块，用于将获取的每个所述SMART数据的S个梯度数据集合作为S个新的第一属性数据子集合分别加入所述每个SMART数据。

8.如权利要求7所述的装置，其特征在于，所述第二获取模块具体包括：

第一生成单元，用于从每个所述SMART数据中的第s个属性对应的第一属性数据子集合中依次选取M个属性数据，以生成P个第二属性数据子集合，其中，P＝N-M+1，N为所述属性s对应的第一属性数据子集合中属性数据的总数，s＝1…S；

第二生成单元，用于分别计算所述P个第二属性数据子集合对应的P个梯度数据，并根据所述P个梯度数据生成所述属性s对应的梯度数据集合。

9.如权利要求8所述的装置，其特征在于，所述第二生成单元具体包括：

第一获取子单元，用于通过加权最小二乘法分别获取所述P个第二属性数据子集合对应的P个拟合系数，其中，所述P个拟合系数中第i个拟合系数k_i＝(Z-b*Y)/X，

其中，i＝1…P，

$$X=\sum_{j=i}^{M+i-1} w_j*x_j^2,$$

$$Y=\sum_{j=i}^{M+i-1} w_j*x_i,$$

$$Z=\sum_{j=i}^{M+i-1} w_j*x_j*y_j,$$

$$b=(Z*Y-\sum_{j=i}^{M+i-1}(w_j*y_j)*X)/(Y*Y-X*\sum_{j=i}^{M+i-1} w_j),$$

w_j为所述属性s对应的第一属性数据子集合中第j个属性数据对应的预设权重，x_j为所述属性s对应的第一属性数据子集合中第j个属性数据的检测时间，y_j为所述属性s对应的第一属性数据子集合中第j个属性数据；

第二获取子单元，用于通过以下公式分别获取所述P个梯度数据中的第i个梯度数据$Grad_i$：

$Grad_i＝k_i*(M-1)*y_{M+i-1}$。

10.如权利要求7-9任一项所述的装置，其特征在于，所述第一生成模块具体包括：

处理单元，用于通过以下公式对所述Q个SMART数据进行归一化处理：

$g(x)＝sign(x)×\log_y|x|$，

其中，x为所述Q个SMART数据中的一个属性数据，g(x)为属性数据x对应的归一化后的属性数据，所述y可通过以下公式计算：

$y^z \leqslant Value < (y+\Delta y)^z$，

其中，z为预设阈值，Value为所述属性数据x对应的属性的出厂默认值，Δy为预设精度。

11.如权利要求7-9任一项所述的装置，其特征在于，所述修正模块具体包括：

第一获取单元，用于根据所述Q个硬盘类型信息获取与所述Q个硬盘类型信息分别对应的Q个修正值；

修正单元，用于根据与所述Q个硬盘类型信息分别对应的Q个修正值分别对相应的归一化SMART数据进行修正。

12.如权利要求7-9任一项所述的装置，其特征在于，所述修正SMART数据集合中包括与所述S个属性分别对应的S个修正属性数据集合，所述第二生成模块具体包括：

第二获得单元，用于分别获取所述S个属性对应的S个训练特征数据；

排序单元，用于分别对每个训练特征数据中的训练特征值进行排序以生成与所述S个属性对应的S个特征序列(V_i)；

第三获取单元，用于通过以下预设映射规则获取每个属性对应的修正属性数据集合中的每个属性值v的特征值f(v)：

$$\begin{cases} f(v)=V_i & V_i \le v \le [(V_{i+1}-V_i)/2] \\ f(v)=V_i+1 & V_i+[(V_{i+1}-V_i)/2]+1 < v < V_{i+1} \end{cases};$$

第三生成单元，用于根据获取的每个属性对应的修正属性数据集合中的每个属性值的特征值生成所述硬盘特征数据。

硬盘SMART数据中特征数据提取方法和装置

技术领域

[0001] 本发明涉及存储技术领域，特别涉及一种硬盘SMART自我监测、分析和报告技术数据中特征数据提取方法和装置。

背景技术

[0002] 由于硬盘故障可从硬盘SMART(Self Monitoring Analysis And Reporting Technology，自我监测、分析和报告技术)数据中反映出来，因此在硬盘故障预警分析中，可根据硬盘的SMART数据分析硬盘在未来的一段时间之内是否会发生故障。目前，可通过机器学习算法根据SMART数据中的某个属性训练故障预警模型，从而根据该故障预警模型对硬盘的SMART数据进行分析以预测硬盘在未来一段时间内是否能够稳定工作。

[0003] 但是，由于SMART数据中不同属性的特征值表示方式不统一，并且过于离散，很难预测若干不同属性对硬盘的共同影响。而且，有些属性在训练模型的时候，存在特征值缺失的情况，增大了分析SMART数据的难度，使得模型预测不准确。此外，不同厂商的硬盘数据的特征值计算方式不统一，不利于统一的数值特征表示，因此需要对每个厂商的硬盘的SMART数据分别训练故障预警模型以进行故障预警分析，这就需要多次进行模型训练，由此使得分析成本被大大增高。

发明内容

[0004] 本发明旨在至少在一定程度上解决上述技术问题。

[0005] 为此，本发明的第一个目的在于提出一种硬盘SMART数据中特征数据提取方法，该方法无需多个故障预警模型，仅通过同一个故障预警模型即可实现对不同硬盘的故障预警测试和分析，提高了故障预警模型的准确性，降低了模型训练、测试和分析成本。

[0006] 本发明的第二个目的在于提出一种硬盘SMART数据中特征数据提取装置。

[0007] 为达上述目的，本发明第一方面实施例提出了一种硬盘SMART数据中特征数据提取方法，包括以下步骤：获取样本硬盘的SMART数据集合，其中，所述SMART数据集合包括Q个SMART数据和与所述Q个SMART数据分别对应的Q个硬盘类型信息；对所述Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据；根据所述Q个硬盘类型信息分别对所述Q个归一化SMART数据进行修正，以生成修正SMART数据集合；根据所述修正SMART数据集合生成硬盘特征数据。

[0008] 本发明实施例的硬盘SMART数据中特征数据提取方法，通过对样本硬盘的SMART数据进行归一化处理，以及根据硬件类型信息对归一化的硬盘SMART数据进行修正，由此，使硬盘SMART数据具有相同的值域，并且通过对归一化的硬盘SMART数据进行修正分区，从而通过同一个故障预警模型即可实现对不同硬盘的故障预警测试和分析，提高了故障预警模型的准确性，降低了模型训练、测试和分析成本。

[0009] 为达上述目的，本发明第二方面实施例提供了一种硬盘SMART数据中特征数据提取装置，包括：第一获取模块，用于获取样本硬盘的SMART数据集合，其中，所述SMART数据集

合包括Q个SMART数据和与所述Q个SMART数据分别对应的Q个硬盘类型信息；第一生成模块，用于对所述Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据；修正模块，用于根据所述Q个硬盘类型信息分别对所述Q个归一化SMART数据进行修正，以生成修正SMART数据集合；第二生成模块，用于根据所述修正SMART数据集合生成硬盘特征数据。

[0010] 本发明实施例的硬盘SMART数据中特征数据提取装置，通过对样本硬盘的SMART数据进行归一化处理，以及根据硬件类型信息对归一化的硬盘SMART数据进行修正，由此，使硬盘SMART数据具有相同的值域，并且通过对归一化的硬盘SMART数据进行修正分区，从而通过同一个故障预警模型即可实现对不同硬盘的故障预警测试和分析，提高了故障预警模型的准确性，降低了模型训练、测试和分析成本。

[0011] 本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0012] 本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

[0013] 图1是本发明一个实施例的硬盘SMART数据中特征数据提取方法的流程图；

[0014] 图2是本发明另一个实施例的硬盘SMART数据中特征数据提取方法的流程图；

[0015] 图3是本发明一个具体实施例的硬盘SMART数据中梯度数据的归一化分析结果的示意图；

[0016] 图4是本发明的一个具体实施例的硬盘SMART数据中属性数据的归一化分析结果的示意图；

[0017] 图5是本发明一个实施例的硬盘SMART数据中特征数据提取装置的结构示意图；

[0018] 图6是本发明另一个实施例的硬盘SMART数据中特征数据提取装置的结构示意图；以及

[0019] 图7是本发明又一个实施例的硬盘SMART数据中特征数据提取装置的结构示意图。

具体实施方式

[0020] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

[0021] 在本发明的描述中，需要理解的是，术语“中心”、“纵向”、“横向”、“上”、“下”、“前”、“后”、“左”、“右”、“竖直”、“水平”、“顶”、“底”、“内”、“外”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本发明的限制。此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性。

[0022] 在本发明的描述中，需要说明的是，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或一体地连接；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是

两个元件内部的连通。对于本领域的普通技术人员而言，可以具体情况理解上述术语在本发明中的具体含义。

[0023] 目前，可通过机器学习算法根据SMART数据中的某个属性训练故障预警模型，从而预测硬盘在未来一段时间内是否能够稳定工作。然而，现有的故障预警模型无法体现出若干SMART属性对硬盘的共同影响，并且也无法将若干种型号的硬盘SMART数据加到一个故障预警模型当中进行训练。因此，硬盘故障预警并不准确，并且在硬盘预警过程中需要对不同型号的硬盘所对应的不同故障预警模型进行分析，分析成本较高。如果可将不同型号的硬盘SMART数据中的属性值进行处理，使得各个属性值表示统一，则可将若干不同型号的硬件数据加到一个故障预警模型中进行训练，由此可减少故障预警模型训练次数，减少分析成本。为此，本发明提出了一种硬盘SMART数据中特征数据提取方法。

[0024] 图1是本发明一个实施例的硬盘SMART数据中特征数据提取方法的流程图。

[0025] 如图1所示，硬盘SMART数据中特征数据提取方法包括以下步骤

[0026] S11，获取样本硬盘的SMART数据集合。

[0027] 在本发明的一个实施例中，SMART数据集合包括Q个SMART数据和与Q个SMART数据分别对应的Q个硬盘类型信息。其中，SMART数据集合是Q个同种类型和/或不同类型的硬盘SMART中记录的与硬盘相关的例如硬盘寻道出错率、硬盘温度等属性数据，以及与之对应的硬盘类型信息的数据集合。其中，硬盘类型信息是指由硬盘厂商提供的与硬盘相关的例如硬盘型号、硬盘ID(Identity)等数据信息。举例来说，在进行机器算法学习时，硬盘的SMART数据集合中包括多个不同硬盘的硬盘寻道出错率、硬盘加电次数、硬盘温度等属性数据，以及硬盘对应的硬盘型号、硬盘ID(Identity)等信息。

[0028] S12，对Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据。

[0029] 在本发明的一个实施例中，可对Q个不同类型和/或不同类型的硬盘SMART数据中的各个属性数据分别进行归一化处理，从而将SMART数据中具有不同值域的各个属性数据归一化为同一值域内的数据。由此，可实现对不同类型的硬件SMART数据的统一分析和处理。

[0030] S13，根据Q个硬盘类型信息分别对Q个归一化SMART数据进行修正，以生成修正SMART数据集合。

[0031] 为了能够在硬盘故障预警模型的测试结果中分别获取不同硬盘的测试结果，在本发明的一个实施例中，可根据每个硬盘的类型信息对不同的硬盘分别设定相应的数据偏移量，并根据每个硬盘所对应的数据偏移量对Q个归一化SMART数据进行修正以实现对SMART数据集合的分区。

[0032] S14，根据修正SMART数据集合生成硬盘特征数据。

[0033] 本发明实施例的硬盘SMART数据中特征数据提取方法，通过对样本硬盘的SMART数据进行归一化处理，以及根据硬件类型信息对归一化的硬盘SMART数据进行修正，由此，使硬盘SMART数据具有相同的值域，并且通过对归一化的硬盘SMART数据进行修正分区，从而通过同一个故障预警模型即可实现对不同硬盘的故障预警测试和分析，提高了故障预警模型的准确性，降低了模型训练、测试和分析成本。

[0034] 为了使训练出的故障预警模型性能更佳，在对Q个SMART数据进行归一化处理之前，还可通过最小二乘法获得每个SMART数据中的各个属性数据的对应的梯度值。具体地，

图2是本发明另一个实施例的硬盘SMART数据中特征数据提取方法的流程图。

[0035] 如图2所示，硬盘SMART数据中特征数据提取方法包括以下步骤。

[0036] S21，获取样本硬盘的SMART数据集合。

[0037] 在本发明的一个实施例中，SMART数据集合包括Q个SMART数据和与Q个SMART数据分别对应的Q个硬盘类型信息。其中，SMART数据集合是Q个同种类型和/或不同类型的硬盘SMART中记录的与硬盘相关的例如硬盘寻道出错率、硬盘温度等属性数据，以及与之对应的硬盘类型信息的数据集合。其中，硬盘类型信息是指由硬盘厂商提供的与硬盘相关的例如硬盘型号、硬盘ID(Identity)等数据信息。举例来说，在进行机器算法学习时，硬盘的SMART数据集合中包括多个不同硬盘的硬盘寻道出错率、硬盘加电次数、硬盘温度等属性数据，以及硬盘对应的硬盘型号、硬盘ID(Identity)等信息。

[0038] S22，分别获取每个SMART数据中的S个第一属性数据子集合对应的S个梯度数据集合。

[0039] 在本发明的实施例中，SMART数据集合还包括与S个属性分别对应的S个属性数据集合，每个SMART数据中包括与S个属性分别对应的S个第一属性数据子集合。其中，第一属性数据子集合是SMART中的某个属性所对应的数据集合。例如，第一属性数据子集合可以是SMART中的硬盘寻道出错率属性所对应的数据集合。

[0040] 在本发明的实施例中，首先，从每个SMART数据中的第s个属性对应的第一属性数据子集合中依次选取M个属性数据，以生成P个第二属性数据子集合，其中，每个第二属性数据子集合中包括M个属性数据，P＝N-M+1，N为属性s对应的第一属性数据子集合中属性数据的总数，s＝1…S。

[0041] 然后，分别计算P个第二属性数据子集合对应的P个梯度数据，并根据P个梯度数据生成属性s对应的梯度数据集合。具体地，在获得P个第二属性数据子集之后，可通过加权最小二乘法分别获取P个第二属性数据子集合对应的P个拟合系数，具体而言，可先获得P个拟合系数中第i个拟合系数k_i＝(Z-b*Y)/X，其中，i＝1…P，具体地，X、Y、Z和b可以通过下述公式所得：

[0042] $$X=\sum_{j=i}^{M+i-1} w_j*x_j^2,$$

[0043] $$Y=\sum_{j=i}^{M+i-1} w_j*x_i,$$

[0044] $$Z=\sum_{j=i}^{M+i-1} w_j*x_j*y_j,$$

[0045] $$b=(Z*Y-\sum_{j=i}^{M+i-1}(w_j*y_j)*X)/(Y*Y-X*\sum_{j=i}^{M+i-1} w_j),$$

[0046] 其中，w_j为属性s对应的第一属性数据子集合中第j个属性数据对应的预设权重，x_j为属性s对应的第一属性数据子集合中第j个属性数据的检测时间，y_j为属性s对应的第一属性数据子集合中第j个属性数据。

[0047] 在获得P个拟合系数中第i个拟合系数k_i之后，可通过以下公式分别获取P个梯度数据中的第i个梯度数据$Grad_i$：

[0048] $Grad_i = k_i*(M-1)*y_{M+i-1}$

[0049] 其中，$k_i*(M-1)$表示通过加权最小二乘法拟合出来的直线的两个属性数据之间的落差值，落差值的符号表示整条直线的趋势，再与最后的属性数据的y值相乘，可得到最终的梯度值，此时梯度值的大小既可以表示整体变化趋势，还可以表示变化趋势的强度。

[0050] 在获得P个梯度数据之后，可根据P个梯度数据生成属性s对应的梯度数据集合。

[0051] 应当理解，通过上述步骤可以最终获取每个SMART数据中的S个第一属性数据子集合所对应的S个梯度数据集合。

[0052] S23，将获取的每个SMART数据的S个梯度数据集合作为S个新的第一属性数据子集合分别加入每个SMART数据。

[0053] S24，对Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据。

[0054] 在本发明的实施例中，可通过以下公式对Q个SMART数据进行归一化处理：

[0055] $g(x) = sign(x) \times \log_y |x|$，

[0056] 其中，x为Q个SMART数据中的一个属性数据，g(x)为属性数据x对应的归一化后的属性数据，其中，y可通过以下公式计算：

[0057] $y^z \leqslant Value < (y+\Delta y)^z$，

[0058] 其中，z为预设阈值，Value为属性数据x对应的属性的出厂默认值，Δy为预设精度。

[0059] 举例来说，在(1900,2000)所对应的最大梯度值占总数的70％以上时，可通过计算得到的梯度值为Grad＝1.078，y＝1.071，则可得到如图3所示的梯度归一化图像和如图4所示的属性数据归一化图像。

[0060] S25，根据Q个硬盘类型信息分别对Q个归一化SMART数据进行修正，以生成修正SMART数据集合。

[0061] 为了能够在硬盘故障预警模型的测试结果中分别获取不同硬盘的测试结果，在本发明的实施例中，可根据Q个硬盘类型信息获取与Q个硬盘类型信息分别对应的Q个修正值，以及根据与Q个硬盘类型信息分别对应的Q个修正值分别对相应的归一化SMART数据进行修正。举例来说，可根据每个硬盘的类型信息对不同的硬盘分别设定相应的数据偏移量，并根据每个硬盘所对应的数据偏移量对Q个归一化SMART数据进行修正以实现对SMART数据集合的分区。

[0062] S26，根据修正SMART数据集合生成硬盘特征数据。

[0063] 在本发明的实施例中，每个修正SMART数据集合中包括与S个属性分别对应的S个修正属性数据集合。具体地，可分别获取S个属性对应的S个训练特征数据，以及分别对每个训练特征数据中的训练特征值进行排序以生成与S个属性对应的S个特征序列(V_i)。其中，可通过以下预设映射规则获取每个属性对应的修正属性数据集合中的每个属性值v的特征值f(v)：

[0064] $$\begin{cases} f(v)=V_i & V_i \leq v \leq [(V_{i+1}-V_i)/2] \\ f(v)=V_i+1 & V_i+[(V_{i+1}-V_i)/2]+1 < v < V_{i+1} \end{cases}$$

[0065] 在获得每个属性值的特征值之后，可根据获取的每个属性对应的修正属性数据集合中的每个属性值的特征值生成硬盘特征数据。由此，可使得修正SMART数据集合中的每个训练数据所对应的特征值通过映射规则全部应用到故障预警模型中进行训练，避免了训练

模型过程中特征值缺失的缺陷，提高了故障预测模型的准确性。

[0066] 本发明实施例的硬盘SMART数据中特征数据提取方法，在对多个SMART数据进行归一化处理之前，通过最小二乘法获得每个SMART数据中属性数据对应的梯度数据集合，并以梯度数据集合更新对应的第一属性数据集合，由此，可使得SMRAT数据的变化趋势凸显出来，再配合机器学习算法，能够使训练出的故障预警模型性能更佳。

[0067] 为了实现上述实施例，本发明还提出一种硬盘SMART数据中特征数据提取装置。

[0068] 一种硬盘SMART数据中特征数据提取装置，包括：第一获取模块，用于获取样本硬盘的SMART数据集合，其中，SMART数据集合包括Q个SMART数据和与Q个SMART数据分别对应的Q个硬盘类型信息；第一生成模块，用于对Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据；修正模块，用于根据Q个硬盘类型信息分别对Q个归一化SMART数据进行修正，以生成修正SMART数据集合；第二生成模块，用于根据修正SMART数据集合生成硬盘特征数据。

[0069] 图5是本发明一个实施例的硬盘SMART数据中特征数据提取装置的结构示意图。

[0070] 如图5所示，硬盘SMART数据中特征数据提取装置包括：第一获取模块100、第一生成模块200、修正模块300和第二生成模块400。

[0071] 具体地，第一获取模块100用于获取样本硬盘的SMART数据集合。其中，SMART数据集合包括Q个SMART数据和与Q个SMART数据分别对应的Q个硬盘类型信息。换言之，SMART数据集合是Q个同种类型和/或不同类型的硬盘SMART中记录的与硬盘相关的例如硬盘寻道出错率、硬盘温度等属性数据，以及与之对应的硬盘类型信息的数据集合。其中，硬盘类型信息是指由硬盘厂商提供的与硬盘相关的例如硬盘型号、硬盘ID(Identity)等数据信息。举例来说，在进行机器算法学习时，第一获取模块100可获得硬盘的SMART数据集合中的多个不同硬盘的硬盘寻道出错率、硬盘加电次数、硬盘温度等属性数据，以及硬盘对应的硬盘型号、硬盘ID(Identity)等数据信息。

[0072] 第一生成模块200用于对Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据。具体地，在本发明的一个实施例中，第一生成模块200可可对Q个不同类型和/或不同类型的硬盘SMART数据中的各个属性数据分别进行归一化处理，从而将SMART数据中具有不同值域的各个属性数据归一化为同一值域内的数据。由此，可实现对不同类型的硬件SMART数据的统一分析和处理。

[0073] 修正模块300用于根据Q个硬盘类型信息分别对Q个归一化SMART数据进行修正，以生成修正SMART数据集合。具体地，在本发明的一个实施例中，修正模块300可根据每个硬盘的类型信息对不同的硬盘分别设定相应的数据偏移量，并根据每个硬盘所对应的数据偏移量对Q个归一化SMART数据进行修正以实现对SMART数据集合的分区。由此，在硬盘故障预警模型的测试结果中可分别获取不同硬盘的测试结果。

[0074] 第二生成模块400用于根据修正SMART数据集合生成硬盘特征数据。

[0075] 本发明实施例的硬盘SMART数据中特征数据提取装置，通过对样本硬盘的SMART数据进行归一化处理，以及根据硬件类型信息对归一化的硬盘SMART数据进行修正，由此，使硬盘SMART数据能够具有相同的值域，并且通过对归一化的硬盘SMART数据进行修正分区，从而通过同一个故障预警模型即可实现对不同硬盘的故障预警测试和分析，提高了故障预警模型的准确性，降低了模型训练、测试和分析成本。

[0076] 图6是本发明另一个实施例的硬盘SMART数据中特征数据提取装置的结构示意图。

[0077] 如图6所示，硬盘SMART数据中特征数据提取装置包括：第一获取模块100、第一生成模块200、修正模块300、第二生成模块400、第二获取模块500和加入模块600。

[0078] 在本发明的实施例中，SMART数据集合包括与S个属性分别对应的S个属性数据集合，每个SMART数据中包括与S个属性分别对应的S个第一属性数据子集合。其中，第一属性数据子集合是SMART中的某个属性所对应的数据集合。例如，第一属性数据子集合可以是SMART中的硬盘寻道出错率属性所对应的数据集合。

[0079] 具体地，第二获取模块500用于分别获取每个SMART数据中的S个第一属性数据子集合对应的S个梯度数据集合。加入模块600用于将获取的每个SMART数据的S个梯度数据集合作为S个新的第一属性数据子集合分别加入每个SMART数据。

[0080] 图7是本发明又一个实施例的硬盘SMART数据中特征数据提取装置的结构示意图。

[0081] 如图7所示，硬盘SMART数据中特征数据提取装置包括：第一获取模块100、第一生成模块200、修正模块300、第二生成模块400、第二获取模块500和加入模块600。其中，第二获取模块500包括：第一生成单元510和第二生成单元520，其中，第一生成模块200包括处理单元210，修正模块300包括：第一获取单元310和修正单元320，第二生成模块400包括：第二获得单元410、排序单元420和第三获取单元430。其中，第二生成单元520包括：第一获取子单元521和第二获取子单元522。

[0082] 具体地，第一生成单元510用于从每个SMART数据中的第s个属性对应的第一属性数据子集合中依次选取M个属性数据，以生成P个第二属性数据子集合，其中，每个第二属性数据子集合中包括M个属性数据，P＝N-M+1，N为属性s对应的第一属性数据子集合中属性数据的总数，s＝1…S。

[0083] 第二生成单元520用于分别计算P个第二属性数据子集合对应的P个梯度数据，并根据P个梯度数据生成属性s对应的梯度数据集合。

[0084] 处理单元210用于通过以下公式对Q个SMART数据进行归一化处理：

[0085] $g(x)=\mathrm{sign}(x)\times\log_y|x|$，

[0086] 其中，x为Q个SMART数据中的一个属性数据，g(x)为属性数据x对应的归一化后的属性数据，y可通过以下公式计算：

[0087] $y^z\leqslant \mathrm{Value}<(y+\Delta y)^z$，

[0088] 其中，z为预设阈值，Value为属性数据x对应的属性的出厂默认值，Δy为预设精度。

[0089] 第一获取单元310用于根据Q个硬盘类型信息获取与Q个硬盘类型信息分别对应的Q个修正值。

[0090] 修正单元320用于根据与Q个硬盘类型信息分别对应的Q个修正值分别对相应的归一化SMART数据进行修正。

[0091] 第二获得单元410用于分别获取S个属性对应的S个训练特征数据。

[0092] 排序单元420用于分别对每个训练特征数据中的训练特征值进行排序以生成与S个属性对应的S个特征序列(V_i)。

[0093] 第三获取单元430用于通过以下预设映射规则获取每个属性对应的修正属性数据集合中的每个属性值v的特征值f(v)：

[0094] $$\begin{cases} f(v)=V_i & V_i \le v \le [(V_{i+1}-V_i)/2] \\ f(v)=V_i+1 & V_i+[(V_{i+1}-V_i)/2]+1 < v < V_{i+1} \end{cases}$$

[0095] 第三生成单元440用于根据获取的每个属性对应的修正属性数据集合中的每个属性值的特征值生成硬盘特征数据。

[0096] 第一获取子单元521用于通过加权最小二乘法分别获取P个第二属性数据子集合对应的P个拟合系数，其中，P个拟合系数中第i个拟合系数$k_i=(Z-b*Y)/X$，

[0097] 其中，$i=1\cdots P$，

[0098] $$X=\sum_{j=i}^{M+i-1} w_j*x_j^2,$$

[0099] $$Y=\sum_{j=i}^{M+i-1} w_j*x_i,$$

[0100] $$Z=\sum_{j=i}^{M+i-1} w_j*x_j*y_j,$$

[0101] $$b=(Z*Y-\sum_{j=i}^{M+i-1}(w_j*y_j)*X)/(Y*Y-X*\sum_{j=i}^{M+i-1} w_j),$$

[0102] w_j为属性s对应的第一属性数据子集合中第j个属性数据对应的预设权重，x_j为属性s对应的第一属性数据子集合中第j个属性数据的检测时间，y_j为属性s对应的第一属性数据子集合中第j个属性数据。

[0103] 第二获取子单元522用于通过以下公式分别获取P个梯度数据中的第i个梯度数据$Grad_i$：

[0104] $Grad_i=k_i*(M-1)*y_{M+i-1}$。

[0105] 本发明实施例的硬盘SMART数据中特征数据提取装置，通过最小二乘法获得每个SMART数据中属性数据对应的梯度数据集合，并以梯度数据集合更新对应的第一属性数据集合，由此，可使得SMRAT数据的变化趋势凸显出来，再配合机器学习算法，能够使训练出的故障预警模型性能更佳。

[0106] 流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或更多个用于实现特定逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本发明的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本发明的实施例所属技术领域的技术人员所理解。

[0107] 在流程图中表示或在此以其他方式描述的逻辑和/或步骤，例如，可以被认为是用于实现逻辑功能的可执行指令的定序列表，可以具体实现在任何计算机可读介质中，以供指令执行系统、装置或设备(如基于计算机的系统、包括处理器的系统或其他可以从指令执行系统、装置或设备取指令并执行指令的系统)使用，或结合这些指令执行系统、装置或设备而使用。就本说明书而言，"计算机可读介质"可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例(非穷尽性列表)包括以下：具有一个或多个布线的电连接部(电子装置)，便携式计算机盘盒(磁装置)，随机存取存储器(RAM)，只读存储器

(ROM)，可擦除可编辑只读存储器(EPROM或闪速存储器)，光纤装置，以及便携式光盘只读存储器(CDROM)。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为可以例如通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

[0108] 应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列(PGA)，现场可编程门阵列(FPGA)等。

[0109] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

[0110] 此外，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模块中，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模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

[0111] 上述提到的存储介质可以是只读存储器，磁盘或光盘等。

[0112] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

[0113] 尽管已经示出和描述了本发明的实施例，本领域的普通技术人员可以理解：在不脱离本发明的原理和宗旨的情况下可以对这些实施例进行多种变化、修改、替换和变型，本发明的范围由权利要求及其等同限定。

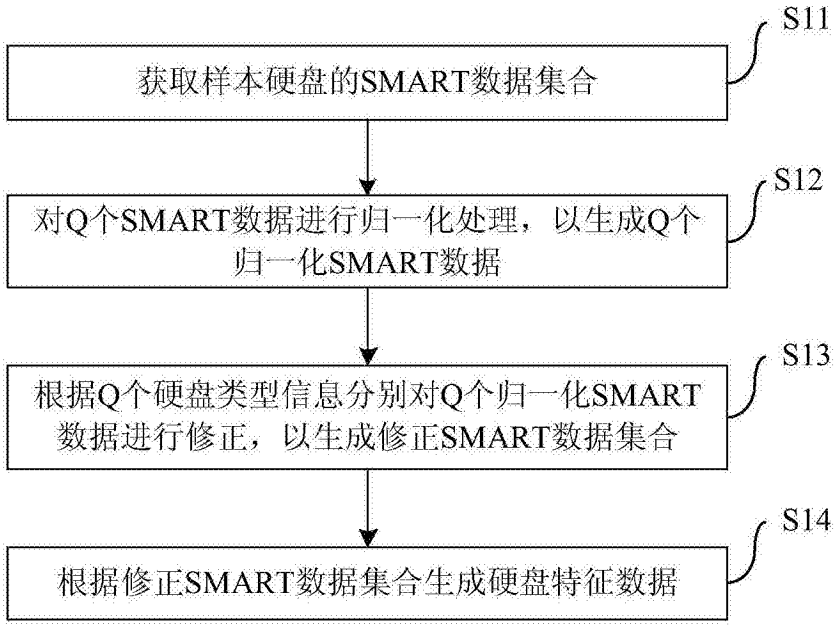

S11
获取样本硬盘的SMART数据集合
S12
对Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据
S13
根据Q个硬盘类型信息分别对Q个归一化SMART数据进行修正，以生成修正SMART数据集合
S14
根据修正SMART数据集合生成硬盘特征数据

图1

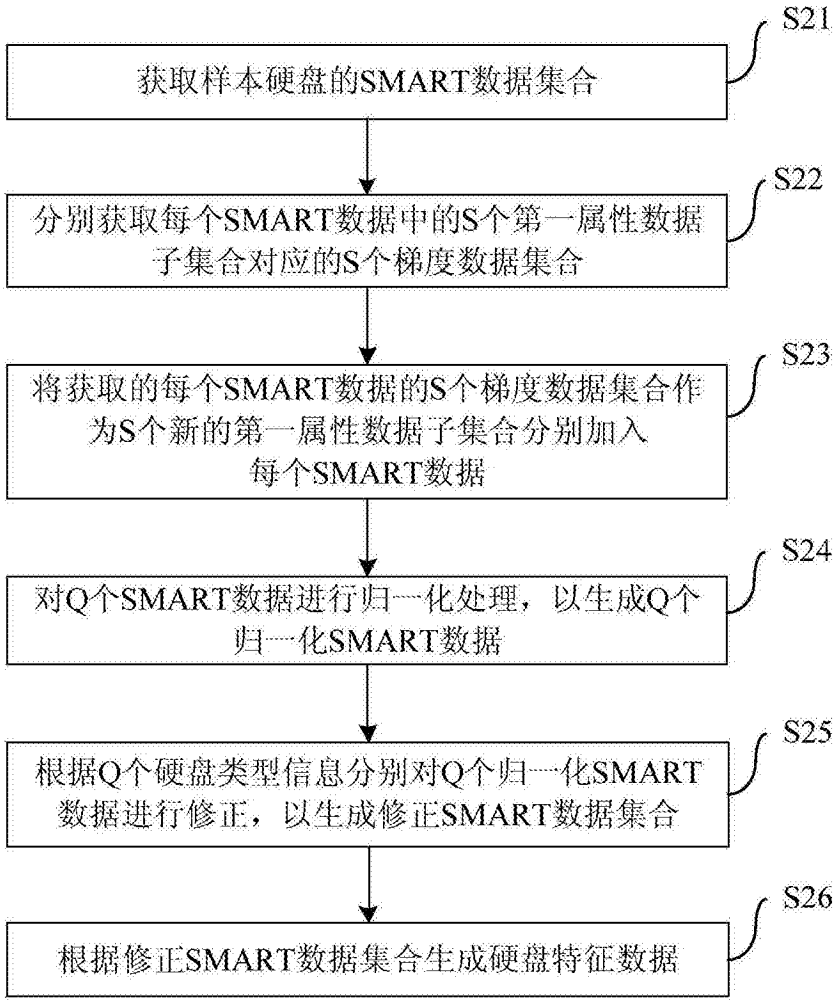

S21
获取样本硬盘的SMART数据集合
S22
分别获取每个SMART数据中的S个第一属性数据子集合对应的S个梯度数据集合
S23
将获取的每个SMART数据的S个梯度数据集合作为S个新的第一属性数据子集合分别加入每个SMART数据
S24
对Q个SMART数据进行归一化处理，以生成Q个归一化SMART数据
S25
根据Q个硬盘类型信息分别对Q个归一化SMART数据进行修正，以生成修正SMART数据集合
S26
根据修正SMART数据集合生成硬盘特征数据

图2

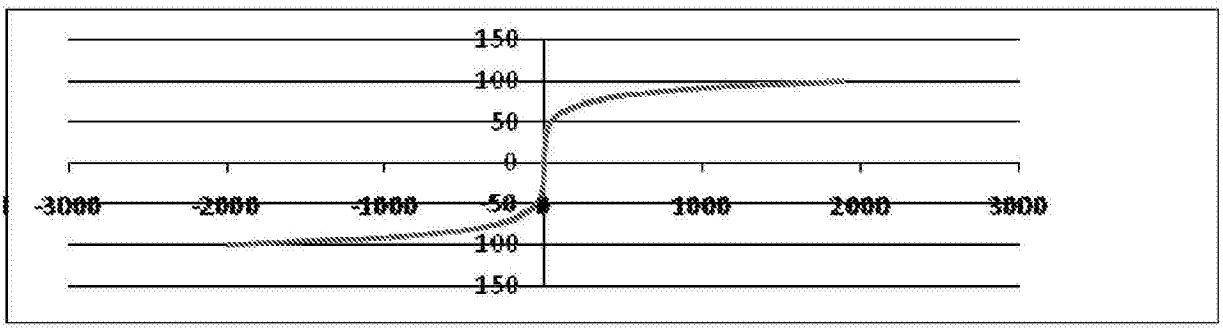

150
100
50
0
-3000
-2000
-1000
-50
1000
2000
3000
-100
-150

图3

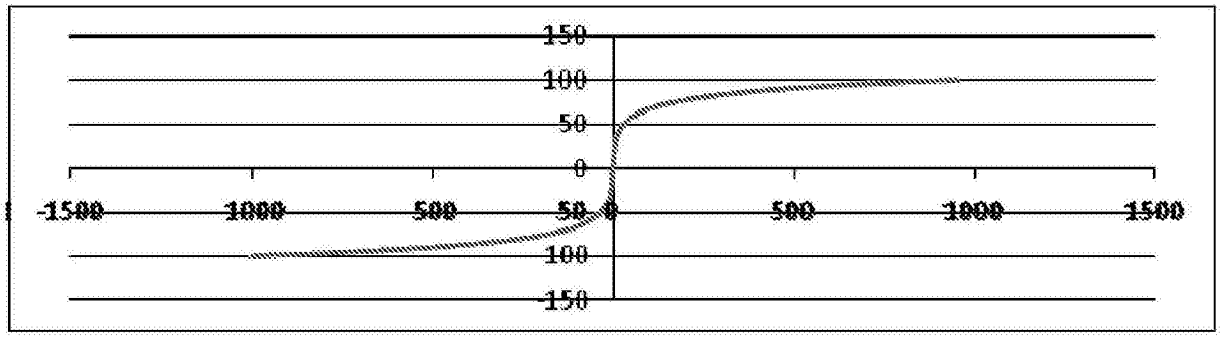

150
100
50
0
-1500
-1000
-500
-50
0
500
1000
1500
100
-150

图4

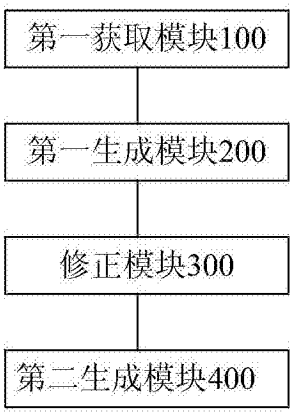

第一获取模块100
第一生成模块200
修正模块300
第二生成模块400

图5

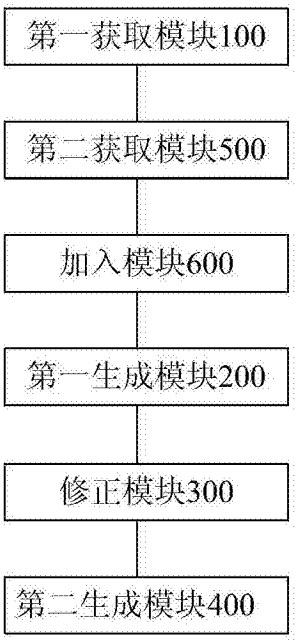

第一获取模块100
第二获取模块500
加入模块600
第一生成模块200
修正模块300
第二生成模块400

图6

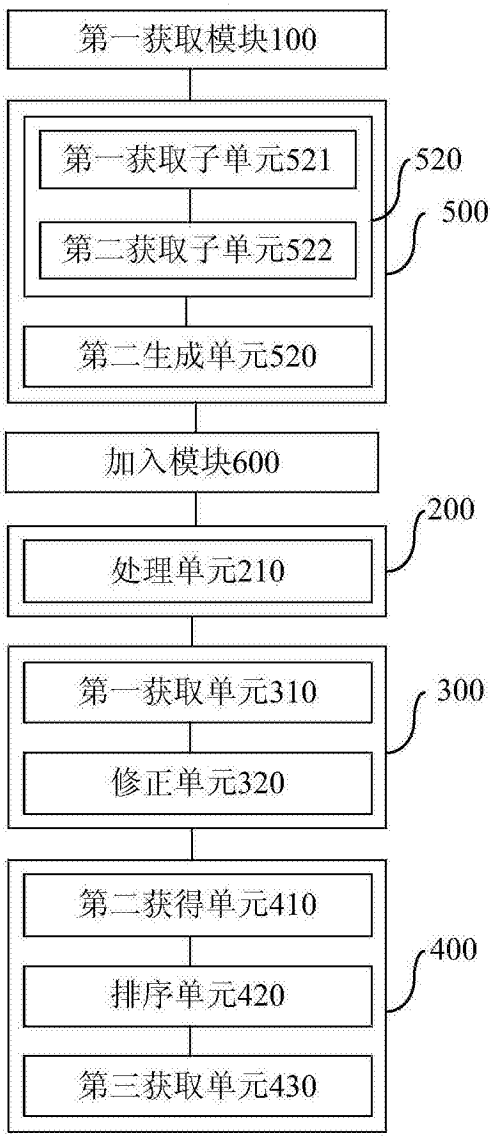

第一获取模块100
第一获取子单元521
第二获取子单元522
520
500
第二生成单元520
加入模块600
处理单元210
200
第一获取单元310
修正单元320
300
第二获得单元410
排序单元420
第三获取单元430
400

图7