

(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
2021 年 12 月 9 日 (09.12.2021)



(10) 国际公布号
WO 2021/244105 A1

(51) 国际专利分类号:
G06K 9/26 (2006.01)

(21) 国际申请号: PCT/CN2021/081443

(22) 国际申请日: 2021 年 3 月 18 日 (18.03.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010495612.8 2020 年 6 月 3 日 (03.06.2020) CN

(71) 申请人: 深信服科技股份有限公司 (SANGFOR TECHNOLOGIES INC.) [CN/CN]; 中国广东省深圳市南山区学苑大道 1001 号南山智园 A1 栋一层, Guangdong 518055 (CN)。

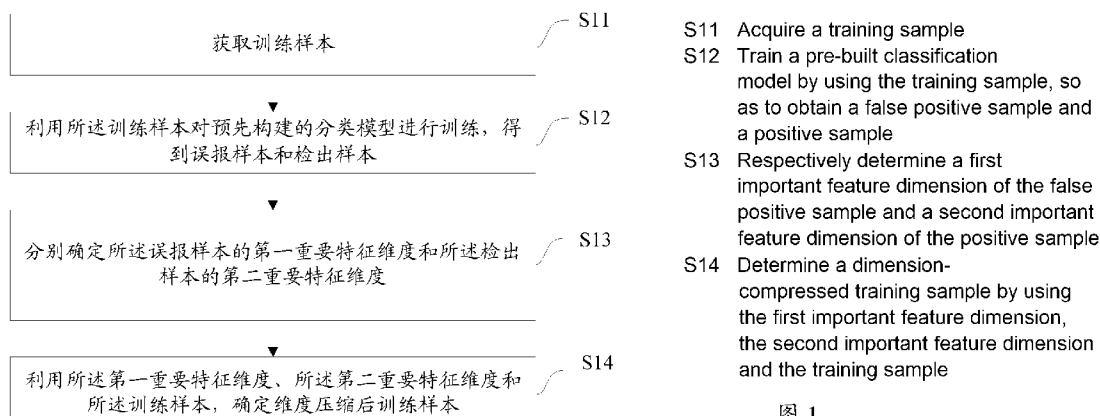
(72) 发明人: 郭开 (GUO, Kai); 中国广东省深圳市南山区学苑大道 1001 号南山智园 A1 栋一层, Guangdong 518055 (CN)。

(74) 代理人: 深圳市深佳知识产权代理事务所 (普通合伙) (SHENPAT INTELLECTUAL PROPERTY AGENCY); 中国广东省深圳市罗湖区南湖街道春风路庐山大厦 B 座 18C2、18D、18E、18E2, Guangdong 518001 (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, IT, JO, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(54) **Title:** FEATURE VECTOR DIMENSION COMPRESSION METHOD AND APPARATUS, AND DEVICE AND MEDIUM

(54) 发明名称: 一种特征向量维度压缩方法、装置、设备、介质



(57) **Abstract:** A feature vector dimension compression method and apparatus, and a device and a medium. The method comprises: acquiring a training sample; training a pre-built classification model by using the training sample, so as to obtain a false positive sample and a positive sample; respectively determining a first important feature dimension of the false positive sample and a second important feature dimension of the positive sample; and determining a dimension-compressed training sample by using the first important feature dimension, the second important feature dimension and the training sample. Therefore, redundant features can be deleted, such that the processing speed of feature vectors is increased. In addition, after an interference feature is deleted, a classification model is trained by using a compressed feature vector, such that the false positive probability of the classification model can be reduced and the importance of a key feature can be improved, thereby improving the sensitivity of the classification model to a black sample and improving the detection rate of malicious files by the classification model.

(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, 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))。

(57) 摘要: 一种特征向量维度压缩方法、装置、设备、介质, 该方法包括: 获取训练样本; 利用所述训练样本对预先构建的分类模型进行训练, 得到误报样本和检出样本; 分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度; 利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本, 确定维度压缩后训练样本。这样能够删除掉冗余的特征, 提高特征向量处理速度, 且在删掉干扰特征之后, 利用压缩后特征向量训练分类模型, 可以减少分类模型的误报概率, 并提升关键特征的重要度, 从而提升分类模型对黑样本的敏感度, 提升分类模型对恶意文件的检出率。

一种特征向量维度压缩方法、装置、设备、介质

本申请要求于 2020 年 06 月 03 日提交中国专利局、申请号为 202010495612.8、发明名称为“一种特征向量维度压缩方法、装置、设备、介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

5 技术领域

本申请涉及信息技术领域，特别涉及一种特征向量维度压缩方法、装置、设备、介质。

背景技术

10 降维就是一种对高维度特征数据进行预处理的方法，是应用非常广泛的数据预处理方法。降维是将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征，从而实现提升数据处理速度的目的。在实际的生产和应用中，降维在一定的信息损失范围内，可以节省大量的时间和成本。现有的一种使用最广泛的数据降维算法是PCA(Principal Component

15 Analysis)，即主成分分析方法，是对主成分进行分析，试图在力保数据信息丢失最少的原则下，对多个变量进行最佳综合简化，即对高维变量空间进行降维处理。另外一种经典的降维方法是LDA (Linear Discriminant Analysis，线性判别分析)，LDA是一种监督学习的降维技术，也就是说数据集的每个样本是有类别输出的，这是和PCA的不同之处。PCA是不考虑

20 样本类别输出的无监督降维技术。但是这两种方法都存在一些问题，首先是解释性差，也即传统降维方法降维后特征不具有可解释性；还有就是性能损耗加大，传统降维方案中存在较多的矩阵运算，不属于轻量级方案，会损失大量的性能；此外大多数降维方案只是在优化特征空间，并没有考虑分类模型的特点，因此降维后的效果并不好。

25

发明内容

有鉴于此，本申请的目的在于提供一种特征向量维度压缩方法、装置、设备、介质，能够删除掉冗余的特征，提高特征向量处理速度，且在删掉

干扰特征之后，利用压缩后特征向量训练分类模型，可以减少分类模型的误报概率，并提升关键特征的重要度，从而提升分类模型对黑样本的敏感度，提升分类模型对恶意文件的检出率，且压缩后特征向量的可解释性强，能够节约性能消耗，提高了降维效果。其具体方案如下：

5 第一方面，本申请公开了一种特征向量维度压缩方法，包括：

获取训练样本；

利用所述训练样本对预先构建的分类模型进行训练，得到误报样本和检出样本；

10 分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度；

利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本。

可选地，所述利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本，包括：

15 将所述训练样本中的第一重要特征维度上的数据删除，并保留所述第二重要特征维度上的数据，得到维度压缩后训练样本。

可选地，所述确定所述误报样本的第一重要特征维度，包括：

利用预设公式分别确定所述误报样本中各个特征维度的目标重要参数，其中，所述目标重要参数用于表示各个特征维度对所述误报样本整体的重要程度；

20 的重要程度；

利用所述目标重要参数确定所述误报样本的第一重要特征维度。

可选地，所述利用预设公式分别确定所述误报样本中各个特征维度的目标重要参数，包括：

25 利用第一预设公式确定所述误报样本中各个样本中的各个特征维度的局部重要参数，其中，所述局部重要参数用于表示各个特征维度对当前样本的重要程度，所述第一预设公式为：

$$f_{m,i} = \sum split(i)$$

其中， $f_{m,i}$ 表示第 m 个样本中的第 i 个特征维度的局部重要参数， $\sum split(i)$ 表示第 m 个样本中的第 i 个特征维度被用来分离决策树的次数；

利用第二预设公式和所述局部重要参数确定所述误报样本中的各个特征维度的目标重要参数，其中，所述第二预设公式为：

$$F_i = \frac{\sum_{m=1}^N f_{m,i}}{N}$$

其中， F_i 表示所述误报样本中的第*i*个特征维度的目标重要参数，*N*表示所述误报样本的数量。

可选地，所述利用所述目标重要参数确定所述误报样本的第一重要特征维度，包括：

将所述误报样本中各个特征维度的目标重要参数与预设重要参数阈值进行比较；

10 如果当前特征维度的目标重要参数大于或等于预设重要参数阈值，则将当前特征维度确定为第一重要特征维度。

可选地，所述利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本之后，还包括：

15 利用所述维度压缩后训练样本训练所述分类模型，得到训练后分类模型。

可选地，所述利用所述维度压缩后训练样本训练所述分类模型，得到训练后分类模型之后，还包括：

获取测试样本；

20 利用所述测试样本对所述训练后分类模型进行评估，直到满足预设条件，得到目标分类模型。

第二方面，本申请公开了一种特征向量维度压缩装置，包括：

样本获取模块，用于获取训练样本；

模型训练模块，用于利用所述训练样本对预先构建的分类模型进行训练，得到误报样本和检出样本；

25 重要特征维度确定模块，用于分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度；

维度压缩模块，用于利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本。

第三方面，本申请公开了一种特征向量维度压缩设备，包括：

存储器和处理器；

其中，所述存储器，用于存储计算机程序；

所述处理器，用于执行所述计算机程序，以实现前述公开的特征向量

5 维度压缩方法。

第四方面，本申请公开了一种计算机可读存储介质，用于保存计算机程序，其中，所述计算机程序被处理器执行时实现前述公开的特征向量维度压缩方法。

可见，本申请先获取训练样本，接着利用所述训练样本对预先构建的
10 分类模型进行训练，得到误报样本和检出样本，然后分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度，再利用所述
第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压
缩后训练样本。由此可见，本申请在获取到训练样本之后，利用所述训练
样本对预先构建的分类模型进行训练，得到误报样本和检出样本。再分别
15 确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维
度，然后利用所述第一重要特征维度、所述第二重要特征维度和所述训练
样本，确定压缩后训练样本，这样能够删除掉冗余的特征，提高特征向量
处理速度，且在删掉干扰特征之后，利用压缩后特征向量训练分类模型，
可以减少分类模型的误报概率，并提升关键特征的重要度，从而提升分类
20 模型对黑样本的敏感度，提升分类模型对恶意文件的检出率，且压缩后特
征向量的可解释性强，能够节约性能消耗，提高了降维效果。

附图说明

为了更清楚地说明本申请实施例或现有技术中的技术方案，下面将对
25 实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，
下面描述中的附图仅仅是本申请的实施例，对于本领域普通技术人员来讲，
在不付出创造性劳动的前提下，还可以根据提供的附图获得其他的附图。

图1为本申请公开的一种特征向量维度压缩方法流程图；

图2为本申请公开的一种具体的特征向量维度压缩方法流程图；

图3为本申请公开的一种具体的特征向量维度压缩方法流程图；

图4为本申请公开的一种特征向量维度压缩装置结构示意图；

图5为本申请公开的一种特征向量维度压缩设备结构图。

5 具体实施方式

下面将结合本申请实施例中的附图，对本申请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申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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，都属于本申请保护的
10 范围。

参见图1所示，本申请实施例公开了一种特征向量维度压缩方法，该方法包括：

步骤 S11：获取训练样本。

在具体的实施过程中，需要先获取训练样本，以便利用获取到的训练
15 样本训练预先构建的分类模型，并根据训练得到的结果确定需要压缩的特征维度。

步骤 S12：利用所述训练样本对预先构建的分类模型进行训练，得到误报样本和检出样本。

可以理解的是，在获取到所述训练样本之后，还需要利用所述训练样
20 本对预先构建的分类模型进行训练，得到误报样本和检出样本。具体的，就是将所述训练样本通过预先构建的分类模型之后，所述分类模型可以将所述样本分为相应的类型，根据分类结果可以确定出检测样本和误报样本。例如，所述样本为 PE 文件(Portable Executable，可移植的可执行的文件)，预先构建的分类模型可以将所述 PE 文件分为正常 PE 文件和病毒 PE 文件，
25 则将分为正常 PE 文件中确实为正常 PE 文件的和分为病毒 PE 文件中确实为病毒 PE 文件的确定为检出样本，将分为正常 PE 文件中实际不为正常 PE 文件的和分为病毒 PE 文件中实际不为病毒 PE 文件的确定为误报样本。其中，所述 PE 文件是微软操作系统上的可执行文件，比如常见的 EXE、DLL、OCX、SYS、COM 都是 PE 文件。在实际应用过程中，所述训练样

本需要先用特征向量进行表示，再将用特征向量表示的训练样本输入到预先构建的分类模型中。特征向量是原始特征对应的特征表达，比如字符串 apple, orange 使用 one-hot 编码后对应的特征向量分别为[1, 0]和[0, 1]。

其中，One-Hot 编码又称为一位有效编码，主要是采用 N 位状态寄存器来对 N 个字符串进行编码，每个字符串都有独立的寄存器位，并且在任意时候只有一位有效，One-Hot 编码是分类变量作为二进制向量的表示，这首先要求将分类值映射到整数值，然后，每个整数值被表示为二进制向量，除了整数的索引之外，它都是零值，它被标记为 1。

5 步骤 S13: 分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度。

在得到所述误报样本和所述检出样本之后，需要分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度。也即分别确定出对于误报样本重要的特征维度和对于检出样本重要的特征维度，这样可以确定出需要压缩的特征维度。

15 步骤 S14: 利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本。

在确定出所述第一重要特征维度、所述第二重要特征维度和所述训练样本之后，便可以利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本确定维度压缩后样本。具体的，就是在确定出所述第一重要特征维度和所述第二重要特征维度之后，相应的可以确定出需要压缩的特征维度，以便进行特征维度压缩。

25 可见，本申请先获取训练样本，接着利用所述训练样本对预先构建的分类模型进行训练，得到误报样本和检出样本，然后分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度，再利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本。由此可见，本申请在获取到训练样本之后，利用所述训练样本对预先构建的分类模型进行训练，得到误报样本和检出样本。再分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度，然后利用所述第一重要特征维度、所述第二重要特征维度和所述训练

样本，确定压缩后训练样本，这样能够删除掉冗余的特征，提高特征向量处理速度，且在删掉干扰特征之后，利用压缩后特征向量训练分类模型，可以减少分类模型的误报概率，并提升关键特征的重要度，从而提升分类模型对黑样本的敏感度，提升分类模型对恶意文件的检出率，且压缩后特征向量的可解释性强，能够节约性能消耗，提高了降维效果。

参见图2所示，本申请实施例公开了一种具体的特征向量维度压缩方法，该方法包括：

步骤S21：获取训练样本。

10 步骤S22：利用所述训练样本对预先构建的分类模型进行训练，得到误报样本和检出样本。

步骤S21和步骤S22的具体实施过程可以参考前述实施例中公开的相应内容，在此不再进行赘述。

15 步骤S23：分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度。

在得到所述误报样本和所述检出样本之后，还需要分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度。具体的，确定所述误报样本的第一重要特征维度，包括：利用预设公式分别确定所述误报样本中各个特征维度的目标重要参数；利用所述目标重要参数确定
20 所述误报样本的第一重要特征维度，其中，所述目标重要参数用于表示各个特征维度对所述误报样本整体的重要程度。

在具体的实施过程中，所述利用预设公式分别确定所述误报样本中各个特征维度的目标重要参数，包括：利用第一预设公式确定所述误报样本中各个样本中的各个特征维度的局部重要参数，其中，所述局部重要参数
25 用于表示各个特征维度对当前样本的重要程度，所述第一预设公式为：

$$f_{m,i} = \sum split(i)$$

其中， $f_{m,i}$ 表示第 m 个样本中的第 i 个特征维度的局部重要参数， $\sum split(i)$ 表示第 m 个样本中的第 i 个特征维度被用来分离决策树的次数；利用第二预设公式和所述局部重要参数确定所述误报样本中的各个特征维度

的目标重要参数，其中，所述第二预设公式为：

$$F_i = \frac{\sum_{m=1}^N f_{m,i}}{N}$$

其中， F_i 表示所述误报样本中的第*i*个特征维度的目标重要参数，*N*表示所述误报样本的数量。同样的，利用预设公式确定所述检出样本的第二重要特征维度，也包括：利用预设公式分别确定所述检出样本中各个特征维度的目标重要参数；利用所述检出样本中各个特征维度的目标重要参数确定所述检出样本的第二重要特征维度。

在所述利用预设公式确定所述误报样本中各个特征维度的目标重要参数之后，还包括利用所述目标重要参数确定所述误报样本的第一重要特征维度。具体的，所述利用所述目标重要参数确定所述误报样本的第一重要特征维度，包括：将所述误报样本中各个特征维度的目标重要参数与预设重要参数阈值进行比较；如果当前特征维度的目标重要参数大于或等于预设重要参数阈值，则将当前特征维度确定为第一重要特征维度。

步骤S24：将所述训练样本中的第一重要特征维度上的数据删除，并保留所述第二重要特征维度上的数据，得到维度压缩后训练样本。

在确定出所述第一重要特征维度和所述第二重要特征维度之后，也即确定出主要影响误报样本的特征维度和主要影响检出样本的特征维度，这样将所述训练样本中的第一重要特征维度上的数据删除，并保留所述第二重要特征维度上的数据，得到维度压缩后训练样本。

在另一种具体的实施方式中，利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本，包括：将所述训练样本中的第一重要特征维度上的数据删除，并保留所述第二重要特征维度上的数据；将所述第二重要特征维度上的数据赋予相应的权重参数，得到维度压缩后训练样本。例如，得到的所述第二重要特征维度为特征维度1、特征维度2、特征维度3，且相应的目标重要参数排序为特征维度1>特征维度2>特征维度3，则将特征维度1上的数据保留，并乘以权重10，特征维度2上的数据保留，并乘以权重6，特征维度3上的数据保留，并乘以权重4，得到维度压缩后训练样本。这样可以进一步扩大对检出样本重要的维

度的影响，提高检出率。

步骤S25：利用所述维度压缩后训练样本训练所述分类模型，得到训练后分类模型。

5 在得到所述训练后训练样本之后，再利用所述压缩后训练样本训练所述分类模型，得到训练后分类模型，这样得到的训练后分类模型性能大大提升，减少分类模型的误报概率，并提升关键特征的重要度，从而提升分类模型对黑样本的敏感度，提升分类模型对恶意文件的检出率。

步骤S26：获取测试样本。

10 在得到所述训练后分类模型之后，还需要对所述训练后分类模型进行评估，具体的，需要先获取测试样本，其中，所述测试样本可以是待分类的PE文件。

步骤S27：利用所述测试样本对所述训练后分类模型进行评估，直到满足预设条件，得到目标分类模型。

15 可以理解的是，在得到所述测试样本之后，还需要利用所述测试样本对所述训练后分类模型进行评估，直到满足预设条件，得到目标分类模型。也即，利用所述测试样本对所述训练后分类模型进行评估，直到所述训练后分类模型的检测性能不再提升，得到目标分类模型。

20 参见图3所示，为一种特征向量维度压缩方法，获取恶意文件，并利用所述恶意文件训练构建的分类模型，确定出误报样本和检出样本，并对误报样本和检出样本进行特征重要性计算，得到误报维度和检出维度，并误报维度删除，并保留检出维度，得到新的训练样本，再利用所述新的训练样本训练所述分类模型，并评估训练后的分类模型的性能是否达标，如果否，则需要在获取恶意文件，如果是，则结束。

25 参见图4所示，本申请实施例公开了一种特征向量维度压缩装置，包括：样本获取模块11，用于获取训练样本；

模型训练模块12，用于利用所述训练样本对预先构建的分类模型进行训练，得到误报样本和检出样本；

重要特征维度确定模块13，用于分别确定所述误报样本的第一重要特

征维度和所述检出样本的第二重要特征维度;

维度压缩模块14, 用于利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本, 确定维度压缩后训练样本。

可见, 本申请先获取训练样本, 接着利用所述训练样本对预先构建的
5 分类模型进行训练, 得到误报样本和检出样本, 然后分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度, 再利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本, 确定维度压缩后训练样本。由此可见, 本申请在获取到训练样本之后, 利用所述训练样本对预先构建的分类模型进行训练, 得到误报样本和检出样本。再分别
10 确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度, 然后利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本, 确定压缩后训练样本, 这样能够删除掉冗余的特征, 提高特征向量处理速度, 且在删掉干扰特征之后, 利用压缩后特征向量训练分类模型, 可以减少分类模型的误报概率, 并提升关键特征的重要度, 从而提升分类
15 模型对黑样本的敏感度, 提升分类模型对恶意文件的检出率, 且压缩后特征向量的可解释性强, 能够节约性能消耗, 提高了降维效果。

进一步的, 参见图5所示, 本申请实施例还公开了一种特征向量维度压缩设备20, 包括: 处理器21和存储器22。

20 其中, 所述存储器22, 用于存储计算机程序; 所述处理器21, 用于执行所述计算机程序, 以实现前述实施例中公开的特征向量维度压缩装置和特征向量维度压缩方法。

所述特征向量维度压缩设备20具体可以包括但不限于平板电脑、笔记本电脑、台式电脑、服务器或服务器集群等。

25 其中, 处理器21可以包括一个或多个处理核心, 比如四核心处理器、八核心处理器等。处理器21可以采用DSP (digital signal processing, 数字信号处理)、FPGA (field-programmable gate array, 现场可编程门阵列)、PLA (programmable logic array, 可编程逻辑阵列) 中的至少一种硬件来实现。处理器21也可以包括主处理器和协处理器, 主处理器是用于对在唤醒状态

下的数据进行处理的处理单元，也称CPU（central processing unit, 中央处理器）；协处理器是用于对在待机状态下的数据进行处理的低功耗处理器。在一些实施例中，处理器21可以集成有GPU（graphics processing unit, 图像处理单元），GPU用于负责显示屏所需要显示的图像的渲染和绘制。一些实施例中，处理器21可以包括AI（artificial intelligence, 人工智能）处理器，该AI处理器用于处理有关机器学习的计算操作。

存储器22可以包括一个或多个计算机可读存储介质，计算机可读存储介质可以是非易失性的。存储器22还可以包括高速随机存取存储器，以及非易失性存储器，比如一个或多个磁盘存储设备、闪存存储设备。本实施例中，存储器22至少用于存储以下计算机程序221，其中，该计算机程序被处理器21加载并执行之后，能够实现前述任一实施例中公开的特征向量维度压缩方法步骤。另外，存储器22所存储的资源还可以包括操作系统222和数据223等，存储方式可以是短暂存储也可以是永久存储。其中，操作系统222可以是Windows、Unix、Linux等。数据223可以包括各种各样的数据。

在一些实施例中，所述特征向量维度压缩设备20还可包括有显示屏23、输入输出接口24、通信接口25、传感器26、电源27以及通信总线28。

本技术领域人员可以理解，图5中示出的结构并不构成对特征向量维度压缩设备20的限定，可以包括比图示更多或更少的组件。

进一步的，本申请实施例还公开了一种计算机可读存储介质，用于保存计算机程序，其中，所述计算机程序被处理器执行时实现前述任一实施例中公开的特征向量维度压缩方法。

其中，关于上述特征向量维度压缩方法的具体过程可以参考前述实施例中公开的相应内容，在此不再进行赘述。

本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其它实施例的不同之处，各个实施例之间相同或相似部分互相参见即可。对于实施例公开的装置而言，由于其与实施例公开的方法相对应，所以描述的比较简单，相关之处参见方法部分说明即可。

结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块，或者二者的结合来实施。软件模块可以置于随机存储器（RAM）、内存、只读存储器（ROM）、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

最后，还需要说明的是，在本文中，诸如第一和第二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者操作区分开来，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，从而使得一系列包含其他要素的过程、方法、物品或者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特征向量维度压缩方法、装置、设备、介质进行了详细介绍，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

1、一种特征向量维度压缩方法，其特征在于，包括：

获取训练样本；

利用所述训练样本对预先构建的分类模型进行训练，得到误报样本和

5 检出样本；

分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度；

利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本。

10 2、根据权利要求 1 所述的特征向量维度压缩方法，其特征在于，所述利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本，包括：

将所述训练样本中的第一重要特征维度上的数据删除，并保留所述第二重要特征维度上的数据，得到维度压缩后训练样本。

15 3、根据权利要求 1 所述的特征向量维度压缩方法，其特征在于，所述确定所述误报样本的第一重要特征维度，包括：

利用预设公式分别确定所述误报样本中各个特征维度的目标重要参数，其中，所述目标重要参数用于表示各个特征维度对所述误报样本整体的重要程度；

20 利用所述目标重要参数确定所述误报样本的第一重要特征维度。

4、根据权利要求 3 所述的特征向量维度压缩方法，其特征在于，所述利用预设公式分别确定所述误报样本中各个特征维度的目标重要参数，包括：

25 利用第一预设公式确定所述误报样本中各个样本中的各个特征维度的局部重要参数，其中，所述局部重要参数用于表示各个特征维度对当前样本的重要程度，所述第一预设公式为：

$$f_{m,i} = \sum split(i)$$

其中， $f_{m,i}$ 表示第 m 个样本中的第 i 个特征维度的局部重要参数， $\sum split(i)$ 表示第 m 个样本中的第 i 个特征维度被用来分离决策树的次数；

利用第二预设公式和所述局部重要参数确定所述误报样本中的各个特征维度的目标重要参数，其中，所述第二预设公式为：

$$F_i = \frac{\sum_{m=1}^N f_{m,i}}{N}$$

5 其中， F_i 表示所述误报样本中的第*i*个特征维度的目标重要参数，*N*表示所述误报样本的数量。

5、根据权利要求3所述的特征向量维度压缩方法，其特征在于，所述利用所述目标重要参数确定所述误报样本的第一重要特征维度，包括：

将所述误报样本中各个特征维度的目标重要参数与预设重要参数阈值进行比较；

10 如果当前特征维度的目标重要参数大于或等于预设重要参数阈值，则将当前特征维度确定为第一重要特征维度。

6、根据权利要求1所述的特征向量维度压缩方法，其特征在于，所述利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本之后，还包括：

15 利用所述维度压缩后训练样本训练所述分类模型，得到训练后分类模型。

7、根据权利要求6所述的特征向量维度压缩方法，其特征在于，所述利用所述维度压缩后训练样本训练所述分类模型，得到训练后分类模型之后，还包括：

20 获取测试样本；

利用所述测试样本对所述训练后分类模型进行评估，直到满足预设条件，得到目标分类模型。

8、一种特征向量维度压缩装置，其特征在于，包括：

样本获取模块，用于获取训练样本；

25 模型训练模块，用于利用所述训练样本对预先构建的分类模型进行训练，得到误报样本和检出样本；

重要特征维度确定模块，用于分别确定所述误报样本的第一重要特征维度和所述检出样本的第二重要特征维度；

维度压缩模块，用于利用所述第一重要特征维度、所述第二重要特征维度和所述训练样本，确定维度压缩后训练样本。

9、一种特征向量维度压缩设备，其特征在于，包括：

存储器和处理器；

5 其中，所述存储器，用于存储计算机程序；

所述处理器，用于执行所述计算机程序，以实现权利要求 1 至 7 任一项所述的特征向量维度压缩方法。

10、一种计算机可读存储介质，其特征在于，用于保存计算机程序，其中，所述计算机程序被处理器执行时实现如权利要求 1 至 7 任一项所述

10 的特征向量维度压缩方法。

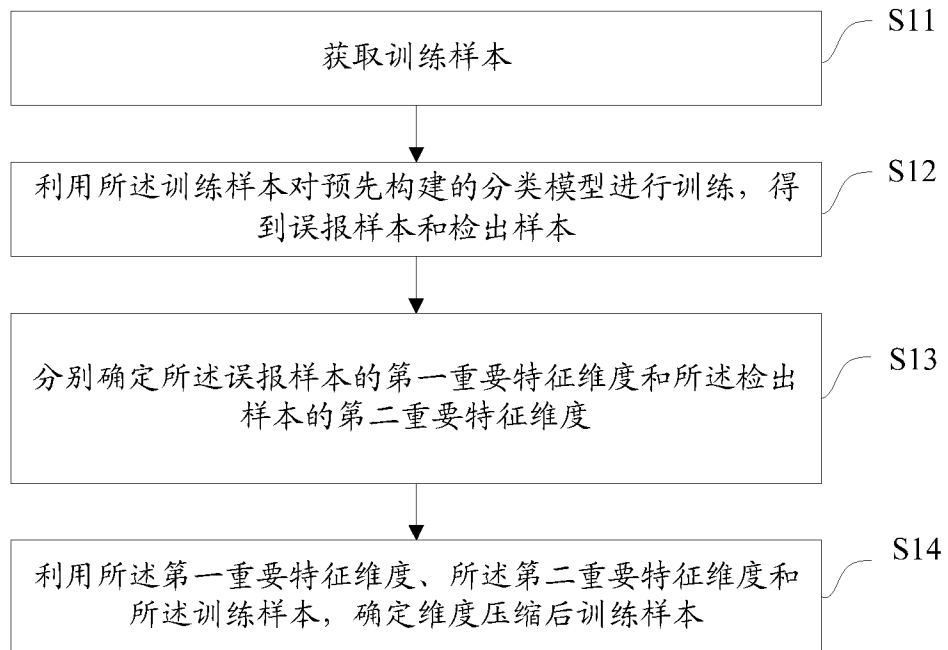


图 1

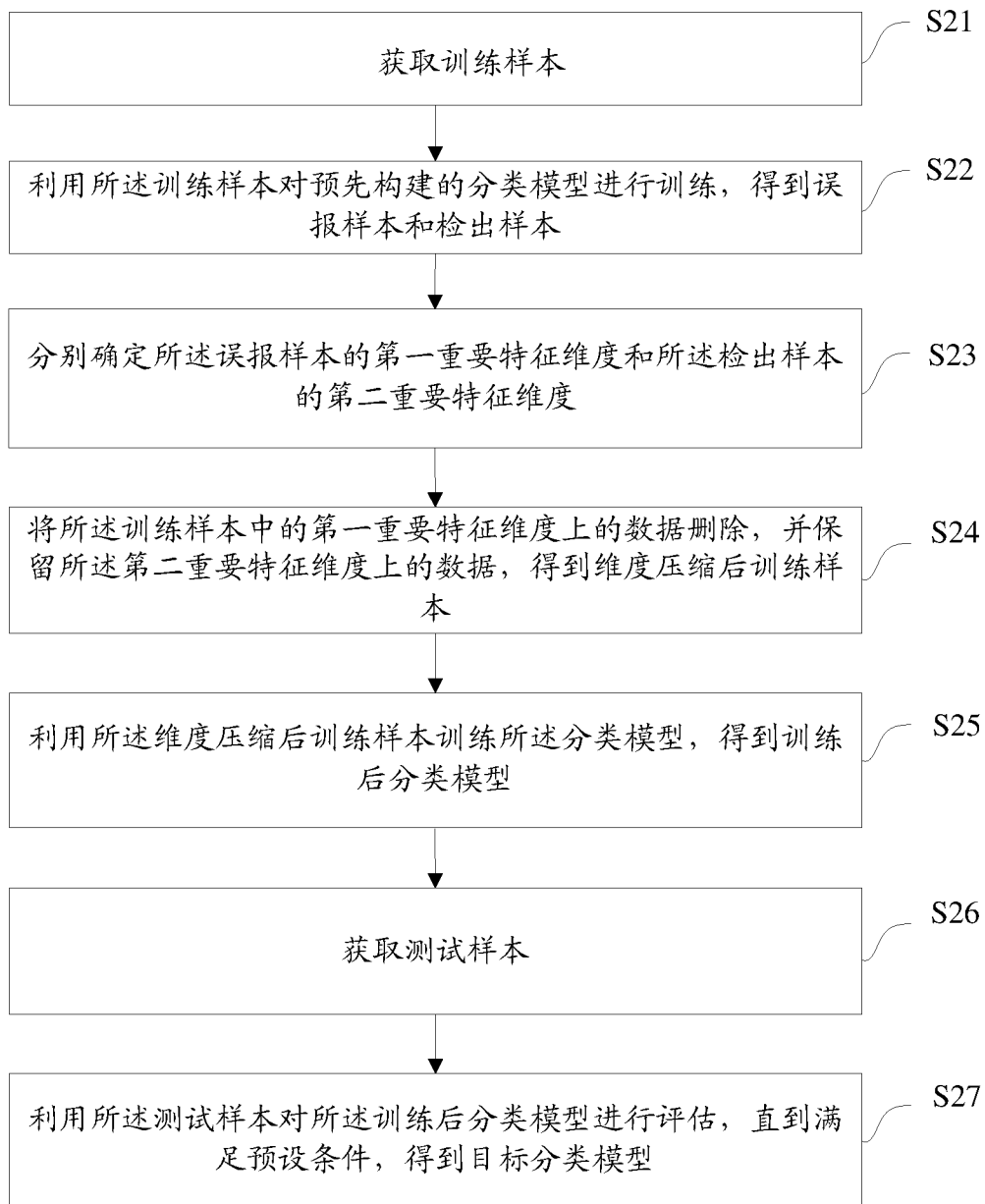


图 2

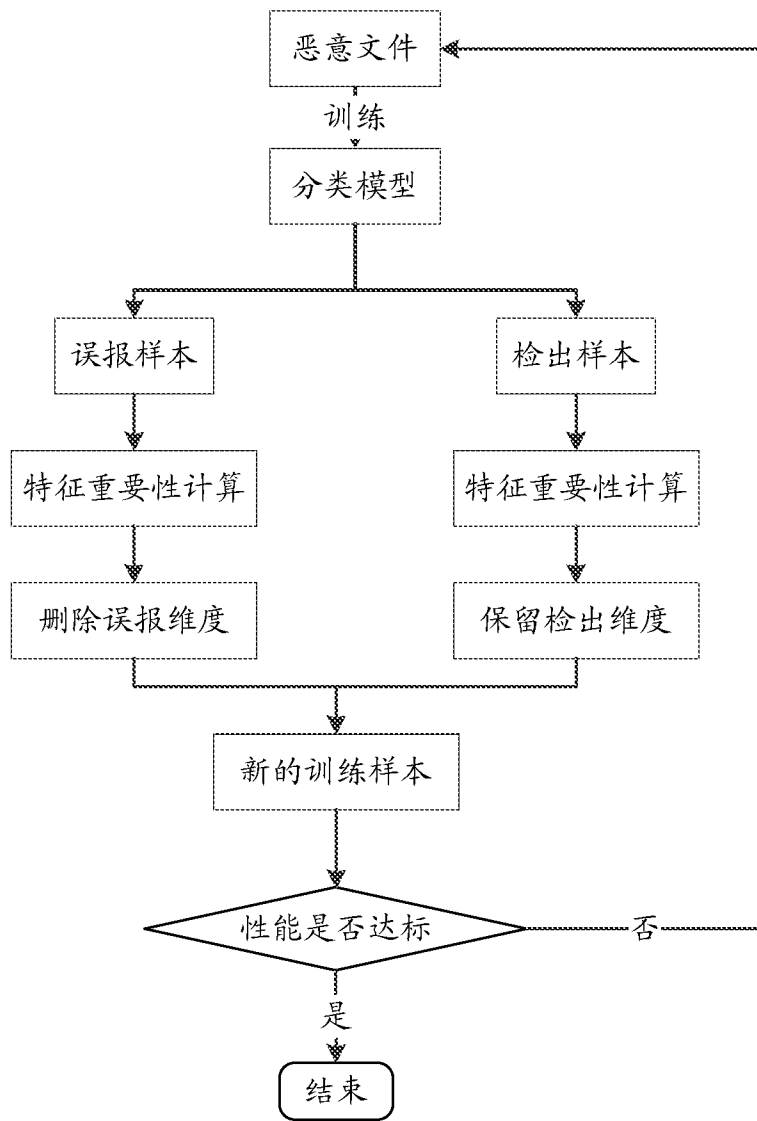


图 3

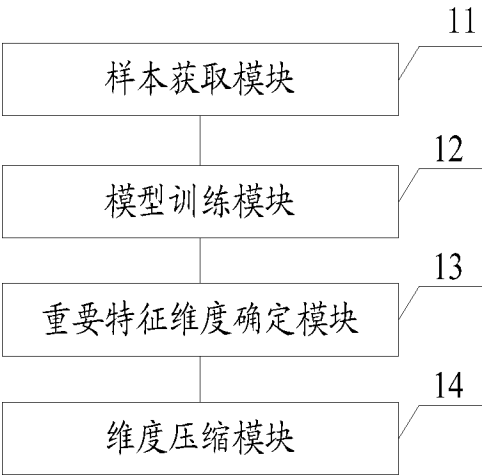


图 4

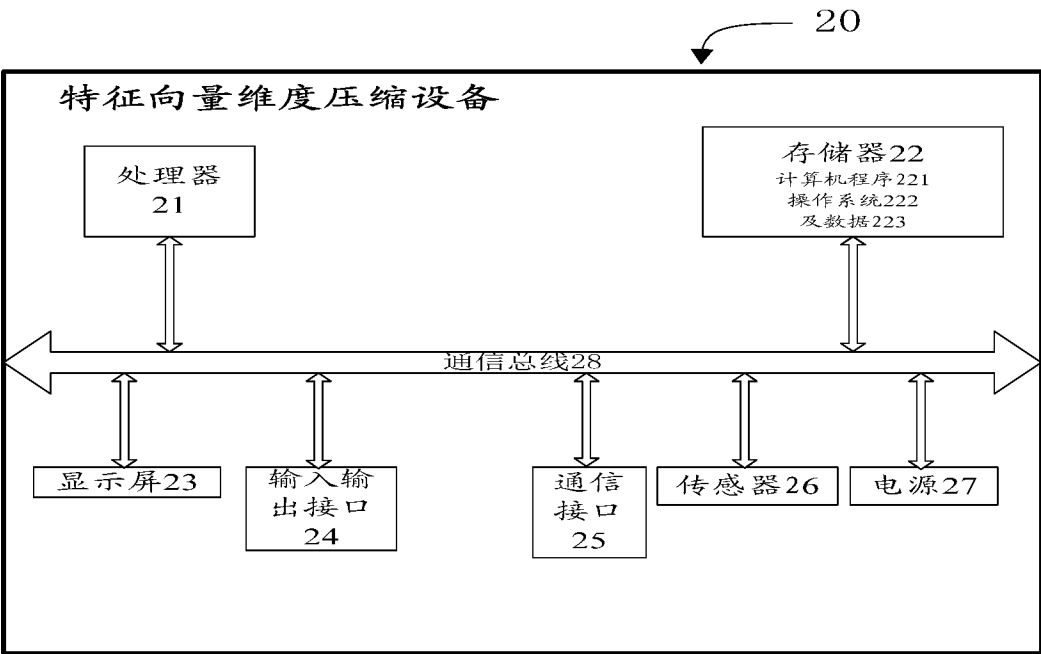


图 5

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/081443

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/26(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)

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)

CNABS, CNTXT, CNKI, VEN, USTXT, EPTXT, WOTXT: 模型, 维度, 降, 样本, model, dimension, reducing, sampling

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 109934167 A (XI'AN JIAOTONG UNIVERSITY) 25 June 2019 (2019-06-25) description, paragraphs 7-67, figure 1	1-10
A	CN 103544499 A (JIANGNAN UNIVERSITY et al.) 29 January 2014 (2014-01-29) description, paragraphs 5-19	1-10
A	CN 102682300 A (TONGJI UNIVERSITY) 19 September 2012 (2012-09-19) entire document	1-10
A	CN 101464907 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES) 24 June 2009 (2009-06-24) entire document	1-10
A	US 2018150724 A1 (CYLANCE INC) 31 May 2018 (2018-05-31) entire document	1-10

☐ 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

24 May 2021

Date of mailing of the international search report

23 June 2021

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088

China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2021/081443

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	109934167	A	25 June 2019	None	
CN	103544499	A	29 January 2014	CN 103544499 B	05 April 2017
CN	102682300	A	19 September 2012	None	
CN	101464907	A	24 June 2009	CN 101464907 B	11 May 2011
US	2018150724	A1	31 May 2018	None	

国际检索报告

国际申请号

PCT/CN2021/081443

A. 主题的分类

G06K 9/26 (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)

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))

CNABS, CNTXT, CNKI, VEN, USTXT, EPTXT, WOTXT: 模型, 维度, 降, 样本, model, dimension, reducing, sampling

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
A	CN 109934167 A (西安交通大学) 2019年 6月 25日 (2019 - 06 - 25) 说明书第7-67段, 图1	1-10
A	CN 103544499 A (江南大学 等) 2014年 1月 29日 (2014 - 01 - 29) 说明书第5-19段	1-10
A	CN 102682300 A (同济大学) 2012年 9月 19日 (2012 - 09 - 19) 全文	1-10
A	CN 101464907 A (中国科学院计算技术研究所) 2009年 6月 24日 (2009 - 06 - 24) 全文	1-10
A	US 2018150724 A1 (CYLANCE INC) 2018年 5月 31日 (2018 - 05 - 31) 全文	1-10

☐ 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

2021年 5月 24日

国际检索报告邮寄日期

2021年 6月 23日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (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

胡嫵

电话号码 86-(20)-28950387

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/081443

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	109934167	A	2019年 6月 25日	无	
CN	103544499	A	2014年 1月 29日	CN 103544499	B 2017年 4月 5日
CN	102682300	A	2012年 9月 19日	无	
CN	101464907	A	2009年 6月 24日	CN 101464907	B 2011年 5月 11日
US	2018150724	A1	2018年 5月 31日	无	