

(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 31 日 (31.08.2017)



(10) 国际公布号
WO 2017/143914 A1

- (51) 国际专利分类号:
G06F 15/18 (2006.01)
- (21) 国际申请号: PCT/CN2017/073340
- (22) 国际申请日: 2017 年 2 月 13 日 (13.02.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610105840.3 2016 年 2 月 25 日 (25.02.2016) CN
- (71) 申请人: 阿里巴巴集团控股有限公司 (ALIBABA GROUP HOLDING LIMITED) [—/CN]; 开曼群岛英属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: 代斌 (DAI, Bin); 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)。李岫 (LI, Shen); 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)。姜晓燕 (JI-ANG, Xiaoyan); 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); 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)。漆远 (QI, Yuan); 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)。褚崴 (CHU, Wei); 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); 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)。付子豪 (FU, Zihao); 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: 北京三友知识产权代理有限公司 (BEIJING SANYOU INTELLECTUAL PROPERTY AGENCY LTD.); 中国北京市金融街 35 号国际企业大厦 A 座 16 层, Beijing 100033 (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,

[见续页]

(54) Title: METHOD FOR TRAINING MODEL USING TRAINING DATA, AND TRAINING SYSTEM

(54) 发明名称: 一种利用训练数据训练模型的方法和训练系统

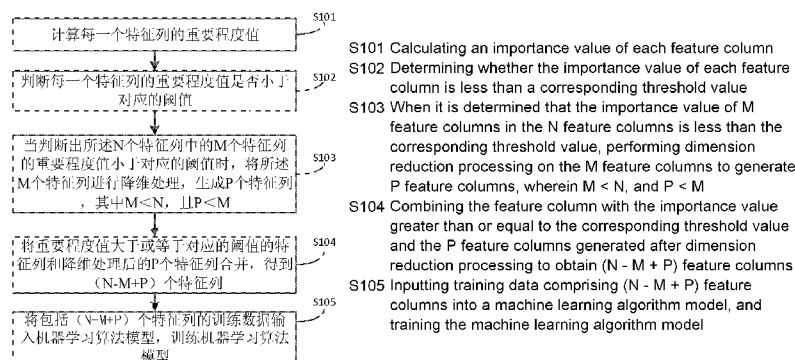


图2

(57) Abstract: A method for training a model using training data, and a training system. The training data comprises a plurality of samples, wherein each sample comprises N features, and the features in each sample constitute N feature columns. The method comprises: calculating an importance value of each feature column (S101); determining whether the importance value of each feature column is less than a corresponding threshold value (S102); when it is determined that the importance value of M feature columns in the N feature columns is less than the corresponding threshold value, performing dimension reduction processing on the M feature columns to generate P feature columns, wherein $M < N$, and $P < M$ (S103); combining the feature column with the importance value greater than or equal to the corresponding threshold value and the P feature columns generated after dimension reduction processing to obtain $(N - M + P)$ feature columns (S104); and inputting training data comprising $(N - M + P)$ feature columns into a machine learning algorithm model, and training the machine learning algorithm model (S105). By means of the present method and system, an important feature and an auxiliary feature can be processed distinctively, thereby achieving the purposes of being able to control a training feature parameter and improving the accuracy of model training.

(57) 摘要:

[见续页]



WO 2017/143914 A1



SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA,
UG, US, UZ, VC, VN, ZA, ZM, ZW。

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)。

- (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,

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

一种利用训练数据训练模型的方法和训练系统, 所述训练数据包括多个样本, 每个样本包括 N 个特征, 每个样本中的特征构成 N 个特征列, 所述方法包括: 计算每个特征列的重要程度值 (S101); 判断每个特征列的重要程度值是否小于对应的阈值 (S102); 当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时, 将所述 M 个特征列进行降维处理, 生成 P 个特征列, 其中 $M < N$, 且 $P < M$ (S103); 将重要程度值大于或等于对应的阈值的特征列和降维处理后生成的 P 个特征列合并, 得到 $(N-M+P)$ 个特征列 (S104); 将包括 $(N-M+P)$ 个特征列的训练数据输入机器学习算法模型, 训练机器学习算法模型 (S105)。本方法和系统能够对重要特征和辅助特征进行区别处理, 达到训练特征参数可控, 并提高模型训练的准确性的目的。

一种利用训练数据训练模型的方法和训练系统

本申请要求 2016 年 02 月 25 日递交的申请号为 201610105840.3、发明名称为“一种利用训练数据训练模型的方法和训练系统”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

本申请涉及数据处理领域，尤其涉及一种利用训练数据训练模型的方法和训练系统。

10 背景技术

在数据挖掘的业务场景中，经常需要利用大规模的训练数据训练机器学习算法模型，进行分类或者回归计算。只有用大量数据训练该机器学习算法模型之后，才有可能获得预测精度较高的该算法模型，进而才能在后续根据输入的数据进行各种行为的预测。

15 在当前互联网的环境下，训练数据的数据规模经常达到数十亿甚至上千亿，每个训练数据的特征规模随着业务扩展，也会到一个非常大的数量级。

以 CTR（Click-Through Rate）预估为例，CTR 预估是互联网计算广告中的关键环节，预估准确性直接影响公司广告收入。CTR 预估使用特定的算法模型，输入该算法模型进行计算的训练数据可能达到千亿规模，每个训练数据的特征维度也可能达到百亿的
20 规模。如此庞大的规模输入算法模型中，容易造成机器学习算法模型的计算难度增大、效率降低。

为了减少计算难度并提高计算的效率，常规的解决方案是使用并行计算的方式处理训练数据，例如使用 LR（Logistic Regression）模型进行处理。LR 是广义线性模型，可以进行并行化处理，即将原本海量的训练数据分发到 N 个不同的计算处理器（worker）
25 中计算，每个计算处理器处理的数据量为总数据量的 $1/N$ 。

利用上述方式可以对一定数量级的数据进行处理，但是对于百亿特征*千亿数据的规模下，数据总量可能超过 PB 级别，超出了一般计算集群的计算范围。即使可以利用超大规模的计算集群进行计算，得到最终最优结果的时间也非常长，导致运行效率低，满足不了业务的更新需求。

30 另外，在一种相反的情况下，训练数据可能会存在维度过低、数据量过小的问题，

同样影响机器学习模型的训练精度。

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利用训练数据训练模型的方法和训练系统。

为解决上述问题，本申请一实施例公开一种利用训练数据训练模型的方法，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，所述训练方法包括：

计算每一个特征列的重要程度值；

10 判断每一个特征列的重要程度值是否小于对应的阈值；

当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，将所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $M < N$ ，且 $P < M$ ；

将重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列和降维处理后生成的 P 个特征列合并，得到 $(N-M+P)$ 个特征列；以及

15 将包括所述 $(N-M+P)$ 个特征列的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

本申请另一实施例公开一种利用训练数据训练模型的方法，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，所述训练方法包括：

20 计算每一个特征列的重要程度值；

判断每一个特征列的重要程度值是否大于对应的阈值；

当判断出所述 N 个特征列中的 Q 个特征列的重要程度值大于对应的阈值时，将所述 Q 个特征列进行升维处理，生成 R 个特征列，其中 $R > Q$ ，且 $Q < N$ ；

25 将升维处理后生成的 R 个特征列和重要程度值小于或等于对应的阈值的 $(N-Q)$ 个特征列合并，得到 $(R+N-Q)$ 个特征列；以及

将包括所述 $(R+N-Q)$ 个特征列的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

本申请再一实施例公开一种利用训练数据训练模型的系统，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，所述训练系统包括：

第一计算模块，用于计算每一个特征列的重要程度值；

第一判断模块，用于判断每一个特征列的重要程度值是否小于对应的阈值；

第一降维模块，用于当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，将所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $M < N$ ，且 $P < M$ ；

5 第一合并模块，用于将重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列和降维处理后生成的 P 个特征列合并，得到 $(N-M+P)$ 个特征列；以及

第一训练模块，用于将包括所述 $(N-M+P)$ 个特征列的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

10 本申请再另一实施例公开一种利用训练数据训练模型的系统，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，所述训练系统包括：

第二计算模块，用于计算每一个特征列的重要程度值；

第二判断模块，用于判断每一个特征列的重要程度值是否大于对应的阈值；

15 升维模块，用于当判断出所述 N 个特征列中的 Q 个特征列的重要程度值大于对应的阈值时，将所述 Q 个特征列进行升维处理，生成 R 个特征列，其中 $R > Q$ ，且 $Q < N$ ；

第二合并模块，用于将升维处理后生成的 R 个特征列和重要程度值小于或等于对应的阈值的 $(N-Q)$ 个特征列合并，得到 $(R+N-Q)$ 个特征列；以及

第二训练模块，用于将包括所述 $(R+N-Q)$ 个特征列的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

20 本申请实施例至少具有以下优点：

本申请实施例通过判断每一个特征列的重要程度值，对重要特征和辅助特征进行区别处理，采用保留重要特征、将辅助特征进行降维的方式，或者对应地采用将重要特征升维的方式，对于较大规模的训练数据，能够降低原本特征维度较高的训练数据的维度，在将该降维后的训练数据输入模型进行训练时，能够达到训练特征参数可控、训练数据大幅降低、运行效率提升的目的。同时由于本申请实施例的方法中保留了重要特征，而
25 对辅助特征进行降维，最大限度地减小了该模型的预测准确度的降低程度。同时，本申请实施例通过判断每一个特征列的重要程度值，对于较小规模的训练数据，能够升高重要程度值较高的训练数据的维度，能够达到训练特征参数可控、提高模型训练的准确性
30 的问题。

附图说明

图 1 是本申请实施例的训练数据的示意图；

图 2 是本申请第一实施例的利用训练数据训练模型的方法的流程图；

图 3 是本申请实施例的特征列的重要程度值输出的结果的示意图；

5 图 4 是本申请第二实施例的利用训练数据训练模型的方法的流程图；

图 5 是本申请第三实施例的利用训练数据训练模型的方法的流程图；

图 6 是本申请第四实施例的利用训练数据训练模型的系统的方框图；

图 7 是本申请第五实施例的利用训练数据训练模型的系统的方框图；

图 8 是本申请第六实施例的利用训练数据训练模型的系统的方框图。

10

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所获得的所有其他实施例，都属于本申请保

15 护的范围。

15

本申请的核心思想之一在于，提出一种利用训练数据训练模型的方法和系统，该方法或系统中首先进行特征列选择，通过计算每一个特征列的重要程度值并判断每一个特征列的重要程度值是否小于对应的阈值，选择出重要特征列和辅助特征列。之后将多个辅助特征列进行降维处理，最后将重要特征列和降维处理后的辅助特征列合并，并输入

20 机器学习算法模型，对机器学习算法模型进行训练。

20

第一实施例

本申请第一实施例提出一种利用训练数据训练模型的方法，图 1 是本申请一实施例的训练数据的示意图。图 2 是利用训练数据训练模型的方法的流程图。本申请第一实施

25 例提出的利用训练数据训练模型的方法可以用于训练各种机器学习算法模型，特别地，本申请第一实施例提出的利用训练数据训练模型的方法可以应用于具有高维稀疏特征的训练数据。如图 1 所示，所述的训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列。

25

如图 2 所示，本申请第一实施例的利用训练数据训练模型的方法可以包括如下步骤：

30

S101，计算每一个特征列的重要程度值；

在这一步骤中，该重要程度值用于评估该特征列对该模型在训练后的预测准确度的重要程度。举例来说，该重要程度值可以是该特征列的信息值 IV (Information Value, 后称 IV 值) 或信息增益 IG (Information Gain, 后称 IG 值)。

当该重要程度值为该特征列的信息值 IV 时，举例来说，以二分模型为例，可以通过如下公式计算：

$$WoE = \ln(pctlGood / pctlbad) \quad (1)$$

$$MIV = WoE * (pctlGood - pctlbad) \quad (2)$$

$$IV = \sum MIV \quad (3)$$

该二分模型可以为国外广泛使用的信用卡评分模型，该二分模型尤其在信用风险评估以及金融风险控制领域更是得到了广泛的使用，其原理是将模型变量 WoE 编码方式离散化之后运用 logistic 回归模型进行的一种二分类变量的广义线性模型。在该模型中，每一个训练数据即是每一个信用卡用户的信息，该训练数据包括多个特征，例如信用卡用户的年龄、职业、性别等。每一个训练数据对应一个标签值，当该信用卡用户为违约用户时，该标签值记为 1，当该信用卡用户为正常用户时，该标签值记为 0。

在上述公式 (1) 中，WoE 为显著性权重 (Weight of Evidence)，计算 IV 值之前首先需要进行 WoE 值计算。

公式 (1) 中的 pctlGood 和 pctlBad 分别是通过多个训练数据的特征列中的特征以及每一个数据对应的标签值计算得出。其中 pctlGood 表示将某一特征列 (例如“信用卡用户的年龄”) 进行离散处理后，每一个区间内对应的标签值为 1 的信用卡用户个数与全部区间内对应的标签值为 1 的信用卡用户个数，pctlBad 表示每一个区间内对应的标签值为 0 的信用卡用户个数与全部区间内对应的标签值为 0 的信用卡用户个数。

例如，表 1 所示为将信用卡用户的年龄进行离散处理后获得的每个年龄区间对应的 WoE 值。

Age	#Good	#Bad	WoE
0-10	50	200	$=\ln((50/100)/(200/1000))=\ln((50/200)/(100/1000))$
10-18	20	200	$=\ln((20/100)/(200/1000))=\ln((20/200)/(100/1000))$
18-35	5	200	$=\ln((5/100)/(200/1000))=\ln((5/200)/(100/1000))$
35-50	15	200	$=\ln((15/100)/(200/1000))=\ln((15/200)/(100/1000))$
50 以上	10	200	$=\ln((10/100)/(200/1000))=\ln((10/200)/(100/1000))$

汇总	100	1000	
----	-----	------	--

表 1

在完成 WoE 计算之后，通过公式（2）计算出针对每个年龄区间的 MIV 值，再通过公式（3）计算出每个年龄区间的 MIV 值的总和，从而得出该特征列的全部年龄区间的 IV 值。

- 5 当该重要程度值为该特征列的信息增益 IG 时，该信息增益 $\text{Gain}(S, A)$ 被定义为：

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (4)$$

- 在上述公式（4）中，其中 $\text{Entropy}(S)$ 为数据集 S 的熵， $V(A)$ 是某一特征列 A 所有可能值的集合， S_v 是 S 中某一特征列 A 的值为 v 的子集，权值为某一特征列 S_v 的样例占原始样例 S 的比例 $|S_v|/|S|$ ，即 $\text{Gain}(S, A)$ 是由于给定特征列 A 的值而得到的关于目标函数值的信息。

在计算出 IV 值和/或 IG 值之后，可以对计算获得的 IV 值和/或 IG 值进行输出。图 3 所示为输出的结果的示意图，图 3 中，col 列为特征名，IV 为信息值，IG 为信息增益。

- 以上对每一个特征列的 IV 值和 IG 值的计算方法进行了举例描述。本领域技术人员应当可以明确的是，本领域还有其他多种能够用于评估该特征列对该模型在训练后的预测准确度的重要程度的值，并且计算方法也多种多样，本申请并不特别限制该重要程度值的类型和计算方法。

步骤 S102，判断每一个特征列的重要程度值是否小于对应的阈值；

在这一步骤中，举例来说，可以根据步骤 S101 中计算出的每一个特征列的重要程度值，判断该重要程度值是否小于对应的阈值。

- 20 例如，可以根据步骤 S101 中计算出的每一特征列的 IV 值，判断该 IV 值是否小于对应的阈值。特征列的 IV 值对应的阈值例如为第一阈值 B，在这一步骤中，用每一个特征列的 IV 值与该第一阈值 B 进行比较。

- 同样地，可以根据步骤 S101 中计算出的每一特征列的 IG 值，判断该 IG 值是否小于对应的阈值。特征列的 IG 值对应的阈值例如为第二阈值 C，在这一步骤中，用每一个特征列的 IG 值与该第二阈值 C 进行比较。

此外，还可以同时判断每一个特征列的 IV 值和 IG 值是否同时小于对应的阈值。例如，针对每一个特征列，判断该特征列的 IV 值是否小于上述的第一阈值 B，再判断该特征列的 IG 值是否小于上述的第二阈值 C。

步骤 S103, 当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时, 将所述 M 个特征列进行降维处理, 生成 P 个特征列, 其中 $M < N$, 且 $P < M$;

在这一步骤中, 通过信息值和/或信息增益与对应的阈值比较, 可以确定哪些特征列为重要特征, 哪些特征列为辅助特征。即, 哪些特征列对该模型在训练后的预测准确度影响较大, 即可以视为重要特征, 哪些特征列对该模型在训练后的预测准确度影响较小, 即可以视为辅助特征。

如果仅用 IV 值进行比较, 则可以认为 IV 值小于上述的第一阈值 B 的特征列为辅助特征, IV 值大于或等于上述的第一阈值 B 的特征列为重要特征。

如果仅用 IG 值进行比较, 则可以认为 IG 值小于上述的第二阈值 C 的特征列为辅助特征, IG 值大于或等于上述的第二阈值 C 的特征列为重要特征。

如果同时用 IV 值和 IG 值进行比较, 则可以认为 IV 值小于上述的第一阈值 B 的特征列为辅助特征, 或者 IG 值小于上述的第二阈值 C 的特征列为辅助特征, 只有 IV 值大于或等于上述的第一阈值 B, 且 IG 值大于或等于上述的第二阈值 C 的特征列才可以认为是重要特征。

以下通过一个示例, 描述将所述 M 个信息值和/或信息增益小于对应阈值的特征列进行降维处理的步骤。

某个训练数据包括三个样本, 每个样本包括 N 个特征, 三个样本分别为:

A1 ($a_{11}, a_{12}, a_{13}, \dots, a_{1N}$)

A2 ($a_{21}, a_{22}, a_{23}, \dots, a_{2N}$)

A3 ($a_{31}, a_{32}, a_{33}, \dots, a_{3N}$)

在步骤 S101 中计算每一个特征列的重要程度值的步骤, 例如是计算 a_{11}, a_{21}, a_{31} 构成的特征列的信息值和/或信息增益、 a_{12}, a_{22}, a_{32} 构成的特征列的信息值和/或信息增益、 a_{13}, a_{23}, a_{33} 构成的特征列的信息值和/或信息增益, 直到 a_{1N}, a_{2N}, a_{3N} 构成的特征列的信息值和/或信息增益。

在步骤 S103 中, 根据步骤 S102 中得出的每一个特征列的重要程度值是否小于对应的阈值, 将多个特征列分为重要特征和辅助特征。在本示例中, 重要特征例如是 (a_{11}, a_{21}, a_{31}) 构成的特征列, 以及 (a_{12}, a_{22}, a_{32}) 构成的特征列...直到 ($a_{1(N-M)}, a_{2(N-M)}, a_{3(N-M)}$) 构成的特征列; 辅助特征是后 M 列即 ($a_{1(N-M+1)}, a_{2(N-M+1)}, a_{3(N-M+1)}$) 构成的特征列...直到 (a_{1N}, a_{2N}, a_{3N}) 构成的特征列。

在步骤 S103 中, 可以将 M 个重要程度值小于阈值得辅助特征进行降维处理, 转换

成 P 个特征列。即，原先的 $(a_{1(N-M+1)}, a_{1(N-M+2)} \cdots a_{1N})$ 、 $(a_{2(N-M+1)}, a_{2(N-M+2)} \cdots a_{2N})$ 、 $(a_{3(N-M+1)}, a_{3(N-M+2)} \cdots a_{3N})$ 可以分别经过降维处理，转换为 $(a'_{1(N-M+1)}, a'_{1(N-M+2)} \cdots a'_{1(N-M+P)})$ 、 $(a'_{2(N-M+1)}, a'_{2(N-M+2)} \cdots a'_{2(N-M+P)})$ 、 $(a'_{3(N-M+1)}, a'_{3(N-M+2)} \cdots a'_{3(N-M+P)})$ 。

这一步骤 S103 例如可以通过 minwise hashing 算法实现。minwise hashing 算法广泛的应用于海量数据下的信息检索，在这一算法中，将 $(b=64)$ 位缩小到 b 位，降低了存储空间和计算时间。

该算法的实现例如为：

Algorithm 1 The b -bit minwise hashing algorithm, applied to estimating pairwise resemblances in a collection of N sets.

Input: Sets $S_n \in \Omega = \{0, 1, \dots, D-1\}$, $n = 1$ to N .

Pre-processing:

10

1): Generate k random permutations $\pi_j : \Omega \rightarrow \Omega$, $j = 1$ to k .

2): For each set S_n and each permutation π_j , store the lowest b bits of $\min(\pi_j(S_n))$, denoted by $e_{n,i,j}$, $i = 1$ to b .

Estimation: (Use two sets S_1 and S_2 as an example.)

1): Compute $\hat{E}_b = \frac{1}{k} \sum_{j=1}^k \left\{ \prod_{i=1}^b 1\{e_{1,i,\pi_j} = e_{2,i,\pi_j}\} = 1 \right\}$.

2): Estimate the resemblance by $\hat{R}_b = \frac{\hat{E}_b - C_{1,b}}{1 - C_{2,b}}$, where $C_{1,b}$ and $C_{2,b}$ are from Theorem 1 in Sec. 2.

通过 minwise hashing 算法后，原先辅助特征被转化为一个 $k \times 2^b$ 维度的向量，其中 k, b 为算法指定参数，即，步骤 S103 中的 $P = k \times 2^b$ 。

另外，正如本领域技术人员所知，步骤 S103 并不限于通过 minwise hashing 算法完成，同样可以使用主成分分析算法 (PCA)、Linear Discriminant Analysis (LDA)、局部线性嵌入 (LLE)、拉普拉斯特征映射 (Laplacian Eigenmaps) 等算法，在此并不特别限定。

步骤 S104，将重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列和降维处理后生成的 P 个特征列合并，得到 $(N-M+P)$ 个特征列；

在这一步骤中，可以将步骤 S102 中判断出的重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列与降维处理后生成的 P 个特征列合并，得到包括 $(N-M+P)$ 个特征列的新的训练数据。

举例来说，重要程度大于或等于对应阈值的 $(N-M)$ 个特征列例如为：

$A1', (a_{11}, a_{12}, \dots, a_{1(N-M)})$

$$A2' = (a_{21}, a_{22} \cdots a_{2(N-M)})$$

$$A3' = (a_{31}, a_{32} \cdots a_{3(N-M)}),$$

上述降维处理后生成的辅助特征为:

$$A1' = (a'_{1(N-M+1)}, a'_{1(N-M+2)} \cdots a'_{1(N-M+P)})$$

$$A2' = (a'_{2(N-M+1)}, a'_{2(N-M+2)} \cdots a'_{2(N-M+P)})$$

$$A3' = (a'_{3(N-M+1)}, a'_{3(N-M+2)} \cdots a'_{3(N-M+P)})$$

将上述重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列和降维处理后生成的 P 个特征列合并的结果为:

$$A1' = (a_{11}, a_{12}, \cdots a_{1(N-M)}, \cdots a_{1(N-M+P)})$$

$$A2' = (a_{21}, a_{22}, \cdots a_{2(N-M)}, \cdots a_{2(N-M+P)})$$

$$A3' = (a_{31}, a_{32}, \cdots a_{3(N-M)}, \cdots a_{3(N-M+P)})$$

上述合并后生成的训练数据可以称为特征指纹。合并后生成的训练数据的特征维度为 $(N-M+P)$ 个, 短于原先的 N 个。以训练数据包括 1 亿特征列为例, 取 $k=200$, $b=12$, 重要特征取前 10000 个特征列, 所生成特征列为

2000*2¹²+10000=829100 个, 降维比约为 829100/1000000000=0.008292, 特征和数据不到原先的 1%。

S105, 将包括所述 $(N-M+P)$ 个特征列的训练数据输入机器学习算法模型, 训练所述机器学习算法模型。

在这一步骤中, 将合并后的多个样本输入机器学习算法模型, 对机器学习算法模型进行训练, 或者使用机器学习算法模型进行预测。机器学习算法模型可以是分类算法模型或回归算法模型, 例如逻辑回归 (LR) 算法模型、随机森林算法模型、Gradient Boost Decision Tree 算法模型等等, 本申请并不以此为限。

根据本申请第一实施例提出的利用训练数据训练模型的方法, 通过判断每一个特征列的信息值和/或信息增益, 采用保留重要特征、将辅助特征进行降维的方式, 降低原本特征维度较高的训练数据的维度, 在将该降维后的训练数据输入模型进行训练时, 能够达到训练特征参数可控、训练数据大幅降低、运行效率提升的目的。同时由于本申请实施例的方法中保留了重要特征, 而对辅助特征进行降维, 最大限度地减小了该模型的预测准确度的降低程度。本申请提出的利用训练数据训练模型的方法应用在高维稀疏特征的训练数据中的效果尤其明显, 例如在广告 CTR 数据中, 通过上述的改动, 在广告 CTR 的 1 亿特征数据上, AUC 较全量训练提升 2%, 数据大小为原先数据量 1%。

第二实施例

本申请第二实施例提出一种利用训练数据训练模型的方法，图 4 所示为本申请第二
5 实施例的利用训练数据训练模型的方法的流程图。本申请第二实施例提出的利用训练数
据训练模型的方法可以用于训练各种机器学习算法模型，特别地，本申请提出的利用训
练数据训练模型的方法可以应用于具有高维稀疏特征的训练数据。高维稀疏矩阵，例如
总维度与有数值的维度的比例为 100:1 或者 1000:1 以上。所述的训练数据包括多个样
本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列。

10 如图 4 所示，本申请第二实施例的利用训练数据训练模型的方法可以包括如下步骤：
S201，计算每一个特征列的信息值和/或信息增益；

在这一步骤中，该信息值(Information Value, 后称 IV 值)或信息增益(Information
Gain, 后称 IG 值)均能够评估该特征列对该模型在训练后的预测准确度的重要程度。

15 当该重要程度值为该特征列的信息值 IV 时，举例来说，以二分模型为例，可以通
过步骤 S101 中列出的公式 (1) 至公式 (3) 计算，通过上述方法计算出 IV 值和/或 IG
值之后，可以对计算获得的 IV 值和/或 IG 值进行输出。图 3 所示为输出的结果的示意
图，图 3 中，col 列为特征名，IV 为信息值，IG 为信息增益。

步骤 S202，判断每一个特征列的信息值和/或信息增益是否小于对应的阈值；

20 在这一步骤中，举例来说，可以根据步骤 S201 中计算出的每一个特征列的信息值
和/或信息增益，判断该信息值和/或信息增益是否小于对应的阈值。

例如，可以根据步骤 S201 中计算出的每一特征列的 IV 值，判断该 IV 值是否小于
对应的阈值。特征列的 IV 值对应的阈值例如为第一阈值 B，在这一步骤中，用每一个特
征列的 IV 值与该第一阈值 B 进行比较。

25 同样地，可以根据步骤 S201 中计算出的每一特征列的 IG 值，判断该 IG 值是否小
于对应的阈值。特征列的 IG 值对应的阈值例如为第二阈值 C，在这一步骤中，用每一个
特征列的 IG 值与该第二阈值 C 进行比较。

此外，还可以同时判断每一个特征列的 IV 值和 IG 值是否同时小于对应的阈值。例
如，针对每一个特征列，判断该特征列的 IV 值是否小于上述的第一阈值 B，再判断该特
征列的 IG 值是否小于上述的第二阈值 C。

30 步骤 S203，当判断出所述 N 个特征列中的 M 个特征列的信息值和/或信息增益小于

对应的阈值时, 将所述 M 个特征列进行降维处理, 生成 P 个特征列, 其中 $M < N$, 且 $P < M$;

在这一步骤中, 通过信息值和/或信息增益与对应的阈值比较, 可以确定哪些特征列为重要特征, 哪些特征列为辅助特征。即, 哪些特征列对该模型在训练后的预测准确度影响较大, 即可以视为重要特征, 哪些特征列对该模型在训练后的预测准确度影响较小, 即可以视为辅助特征。

如果仅用 IV 值进行比较, 则可以认为 IV 值小于上述的第一阈值 B 的特征列为辅助特征, IV 值大于或等于上述的第一阈值 B 的特征列为重要特征。

如果仅用 IG 值进行比较, 则可以认为 IG 值小于上述的第二阈值 C 的特征列为辅助特征, IG 值大于或等于上述的第二阈值 C 的特征列为重要特征。

如果同时用 IV 值和 IG 值进行比较, 则可以认为 IV 值小于上述的第一阈值 B 的特征列为辅助特征, 或者 IG 值小于上述的第二阈值 C 的特征列为辅助特征, 只有 IV 值大于或等于上述的第一阈值 B , 且 IG 值大于或等于上述的第二阈值 C 的特征列才可以认为是重要特征。

这一步骤 S203 例如可以通过 minwise hashing 算法实现。minwise hashing 算法广泛的应用于海量数据下的信息检索, 在这一算法中, 将 ($b=64$ 位) 缩小到 b 位, 降低了存储空间和计算时间。

该算法可以使用与第一实施例提出的具体实现方式来实现, 在此不再赘述。

通过 minwise hashing 算法后, 原先辅助特征被转化为一个 $k \times 2^b$ 维度的向量, 其中 k, b 为算法指定参数, 即, 步骤 S203 中的 $P = k \times 2^b$ 。

步骤 S204, 将重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列和降维处理后生成的 P 个特征列合并, 得到 $(N-M+P)$ 个特征列;

在这一步骤中, 可以将步骤 S202 中判断出的信息值和/或信息增益大于或等于对应的阈值的 $(N-M)$ 个特征列与降维处理后生成的 P 个特征列合并, 得到新的训练数据。

举例来说, 信息值和/或信息增益大于或等于对应阈值的 $(N-M)$ 个特征列例如为:

$A1' : (a_{11}, a_{12}, \dots, a_{1(N-M)})$

$A2' : (a_{21}, a_{22}, \dots, a_{2(N-M)})$

$A3' : (a_{31}, a_{32}, \dots, a_{3(N-M)})$,

上述降维处理后生成的辅助特征为:

$A1' : (a'_{1(N-M+1)}, a'_{1(N-M+2)}, \dots, a'_{1(N-M+P)})$

$$A2' \quad ' \quad ' \quad (a'_{2(N-M+1)}, a'_{2(N-M+2)} \cdots a'_{2(N-M+P)})$$

$$A3' \quad ' \quad ' \quad (a'_{3(N-M+1)}, a'_{3(N-M+2)} \cdots a'_{3(N-M+P)})$$

将上述信息值和/或信息增益大于或等于对应的阈值的(N-M)个特征列和降维处理后生成的P个特征列合并的结果为:

$$5 \quad A1' \quad ' \quad ' \quad ' \quad (a_{11}, a_{12}, \cdots a_{1(N-M)}, \cdots a_{1(N-M+P)})$$

$$A2' \quad ' \quad ' \quad ' \quad (a_{21}, a_{22}, \cdots a_{2(N-M)}, \cdots a_{2(N-M+P)})$$

$$A3' \quad ' \quad ' \quad ' \quad (a_{31}, a_{32}, \cdots a_{3(N-M)}, \cdots a_{3(N-M+P)})$$

上述合并后生成的训练数据可以称为特征指纹。合并后生成的训练数据的特征维度为(N-M+P)个,短于原先的N个。以训练数据包括1亿特征列为例,取k=200, b=12, 重要特征取前10000个特征列,所生成特征列为 $2000 \times 2^{12} + 10000 = 829100$ 个,降维比约为 $829100/1000000000 = 0.008292$,特征和数据不到原先的1%。同时由于本申请实施例的方法中保留了重要特征,而对辅助特征进行降维,该特征指纹仍然可以全面反映原先的训练数据,并不会降低训练的准确度。

S205,将包括所述(N-M+P)个特征列的训练数据输入机器学习算法模型,训练所述机器学习算法模型。

在这一步骤中,将合并后的多个样本输入机器学习算法模型,对机器学习算法模型进行训练,或者使用机器学习算法模型进行预测。机器学习算法模型可以是分类算法模型或回归算法模型,例如逻辑回归(LR)算法模型、随机森林算法模型、Gradient Boost Decision Tree 算法模型等等,本申请并不以此为限。

根据本申请第二实施例提出的利用训练数据训练模型的方法,通过判断每一个特征列的信息值和/或信息增益,采用保留重要特征、将辅助特征进行降维的方式,降低原本特征维度较高的训练数据的维度,在将该降维后的训练数据输入模型进行训练时,能够达到训练特征参数可控、训练数据大幅降低、运行效率提升的目的。同时由于本申请实施例的方法中保留了重要特征,而对辅助特征进行降维,该特征指纹仍然可以全面反映原先的训练数据,最大限度地减小了该模型的预测准确度的降低程度。本申请提出的利用训练数据训练模型的方法应用在高维稀疏特征的训练数据中的效果尤其明显,例如在广告CTR数据中,通过上述的改动,在广告CTR的1亿特征数据上,AUC较全量训练提升2%,数据大小为原先数据量1%。

30 第三实施例

本申请第三实施例提出一种利用训练数据训练模型的方法，图 5 所示为本申请第三实施例的利用训练数据训练模型的方法的流程图。本申请第三实施例提出的利用训练数据训练模型的方法可以用于训练各种机器学习算法模型，特别地，本申请提出的利用训练数据训练模型的方法可以应用于具有高维稀疏特征的训练数据。高维稀疏矩阵，例如总维度与有数值的维度的比例为 100:1 或者 1000:1 以上。所述的训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列。

如图 5 所示，本申请第三实施例的利用训练数据训练模型的方法可以包括如下步骤：

S301，计算每一个特征列的重要程度值；

S302，判断每一个特征列的重要程度值是否大于对应的阈值；

10 S303，当判断出所述 N 个特征列中的 Q 个特征列的重要程度值大于对应的阈值时，将所述 Q 个特征列进行升维处理，生成 R 个特征列，其中 $R > Q$ ，且 $Q < N$ ；

S304，将升维处理后生成的 R 个特征列和重要程度值小于或等于对应的阈值的特征列合并；以及

S305，将合并的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

15 其中，步骤 S301 和步骤 S305 可与第二实施例中的步骤 S201 和 S205 相同或相似，在此不再赘述。

在步骤 S302 中，相比于步骤 S102，本实施例中是判断重要程度值是否大于对应的阈值，具体的判断方式与步骤 S102 相似，在此不再赘述。

20 在步骤 S303 中，例如可以参考步骤 S103 中的方法，根据步骤 S302 中得出的每一个特征列的重要程度值是否大于对应的阈值，将多个特征列分为重要特征和辅助特征。在本示例中，重要特征例如是 (a_{11}, a_{21}, a_{31}) 构成的特征列，以及 (a_{12}, a_{22}, a_{32}) 构成的特征列……直到 (a_{1Q}, a_{2Q}, a_{3Q}) 构成的特征列；辅助特征是后 $N-Q$ 列即 $(a_{1(Q+1)}, a_{2(Q+2)}, a_{3(Q+3)})$ 构成的特征列……直到 (a_{1N}, a_{2N}, a_{3N}) 构成的特征列。

25 在步骤 S303 中，可以将 Q 个重要程度值大于阈值的重要特征进行升维处理，转换成 R 个特征列。即，原先的 $(a_{11}, a_{12}, \dots, a_{1Q})$ 、 $(a_{21}, a_{22}, \dots, a_{2Q})$ 、 $(a_{31}, a_{32}, \dots, a_{3Q})$ 可以分别经过升维处理，转换为 $(a'_{11}, a'_{12}, \dots, a'_{1R})$ 、 $(a'_{21}, a'_{22}, \dots, a'_{2R})$ 、 $(a'_{31}, a'_{32}, \dots, a'_{3R})$ 。

30 这一步骤 S303 例如可以通过 minwise hashing 算法实现。minwise hashing 算法广泛的应用于海量数据下的信息检索，在这一算法中，通过参数的设定，将 $(b=64 \text{ 位})$ 扩大到 b 位，提升重要特征的维度。

该算法的实现可以参考步骤 S103 所述，在此不再赘述。

在 S304 中，例如可以将升维处理后生成的 R 个特征列和重要程度值小于或等于对应的阈值的 (N-Q) 个特征列合并，可以参考步骤 S103，在此不再赘述。

另外，在步骤 S303 中，当判断出所述 N 个特征列中的 Q 个特征列的重要程度值大于对应的阈值时，可以利用 minwise Hashing 算法对所述 Q 个特征列进行升维处理，生成 R 个特征列，其中 $R=k*2^b$ ，k 和 b 为算法指定参数。

优选地，上述重要程度值可以为第二实施例中的信息值或信息增益。在此不再赘述。

另外，在步骤 S302 之后，该方法还可以包括：

S302a，当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，
10 将所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $M < N$ ，且 $P < M$ ；

步骤 S304 即将升维处理后生成的 R 个特征列和重要程度值小于或等于对应的阈值的特征列合并的步骤例如可以包括：

将升维处理后生成的 R 个特征列和所述降维处理后生成的 P 个特征列合并，生成 (R+P) 个特征列。

15 在步骤 S305 中，例如，可以将上述合并的 (R+P) 个特征列训练数据输入机器学习算法模型，训练所述机器学习算法模型。

根据本申请第三实施例提出的利用训练数据训练模型的方法，通过判断每一个特征列的重要程度值，采用将重要特征升维的方式，对于较小规模的训练数据，能够升高重要程度值较高的训练数据的维度，提高模型训练的准确性的问题。

20

第四实施例

本申请第四实施例公开一种利用训练数据训练模型的系统，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，如图 6 所示为本申请第四实施例的利用训练数据训练模型的系统以及训练数据的示意图。所述

25 训练系统 400 包括：

第一计算模块 401，用于计算每一个特征列的重要程度值；

第一判断模块 402，用于判断每一个特征列的重要程度值是否小于对应的阈值；

第一降维模块 403，用于当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，将所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $M < N$ ，且 $P < M$ ；
30

第一合并模块 404, 用于将重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列和降维处理后生成的 P 个特征列合并; 以及

第一训练模块 405, 用于将合并后的所述多个样本输入机器学习算法模型, 训练所述机器学习算法模型。

- 5 根据本申请第四实施例提出的利用训练数据训练模型的系统, 通过判断每一个特征列的重要程度值, 采用保留重要特征、将辅助特征进行降维的方式, 降低原本特征维度较高的训练数据的维度, 在将该降维后的训练数据输入模型进行训练时, 能够达到训练特征参数可控、训练数据大幅降低、运行效率提升的目的。同时由于本申请实施例的方法中保留了重要特征, 而对辅助特征进行降维, 降维后的数据能够最大限度地减小了该
- 10 模型的预测准确度的降低程度。

第五实施例

- 本申请第五实施例公开一种利用训练数据训练模型的系统, 所述训练数据包括多个样本, 每一个样本包括 N 个特征, 所述多个样本中对应的特征构成 N 个特征列, 如图 7
- 15 所示为本申请第五实施例的利用训练数据训练模型的系统以及训练数据的示意图。所述训练系统 500 包括:

第一计算模块 501, 用于计算每一个特征列的重要程度值;

第一判断模块 502, 用于判断每一个特征列的重要程度值是否小于对应的阈值;

- 第一降维模块 503, 用于当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时, 将所述 M 个特征列进行降维处理, 生成 P 个特征列, 其中 $M < N$, 且 $P < M$;
- 20

第一合并模块 504, 用于将重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列和降维处理后生成的 P 个特征列合并; 以及

- 第一训练模块 505, 用于将合并后的所述多个样本输入机器学习算法模型, 训练所述机器学习算法模型。
- 25

在一优选实施例中, 所述训练数据还包括对应于每一个样本的标签值, 所述重要程度值为信息值和/或信息增益, 所述计算模块 501 包括:

信息值计算模块 51, 用于利用每一个样本对应的标签值和所述特征列中的多个特征, 计算出所述特征列的信息值; 和/或

- 30 信息增益计算模块 52, 用于利用每一个样本对应的标签值和所述特征列中的多个特

征，计算出所述特征列的信息增益。

在一优选实施例中，所述信息值对应的阈值为第一阈值，所述信息增益对应的阈值为第二阈值。

在一优选实施例中，所述降维模块用于：

- 5 利用 minwise Hashing 算法对所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $P=k*2^b$ ，k 和 b 为算法指定参数。

在一优选实施例中，所述机器学习算法模型为分类算法模型或回归算法模型。

- 10 根据本申请第五实施例提出的利用训练数据训练模型的系统，通过判断每一个特征列的信息值和/或信息增益，采用保留重要特征、将辅助特征进行降维的方式，降低原本特征维度较高的训练数据的维度，在将该降维后的训练数据输入模型进行训练时，能够达到训练特征参数可控、训练数据大幅降低、运行效率提升的目的。同时由于本申请实施例的方法中保留了重要特征，而对辅助特征进行降维，该特征指纹仍然可以全面反映原先的训练数据，最大限度地减小了该模型的预测准确度的降低程度。本申请提出的
- 15 在广告 CTR 数据中，通过上述的改动，在广告 CTR 的 1 亿特征数据上，AUC 较全量训练提升 2%，数据大小为原先数据量 1%。

第六实施例

- 20 本申请第六实施例公开一种利用训练数据训练模型的系统，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，如图 8 所示为本申请第六实施例的利用训练数据训练模型的系统以及训练数据的示意图。所述训练系统 600 包括：

第二计算模块 601，用于计算每一个特征列的重要程度值；

第二判断模块 602，用于判断每一个特征列的重要程度值是否小于对应的阈值；

- 25 升维模块 603，用于当判断出所述 N 个特征列中的 Q 个特征列的重要程度值大于对应的阈值时，将所述 Q 个特征列进行升维处理，生成 R 个特征列，其中 $R>Q$ ，且 $Q<N$ ；

第二合并模块 604，用于将升维处理后生成的 R 个特征列和重要程度值小于或等于对应的阈值的 $(N-Q)$ 个特征列合并，得到 $(R+N-Q)$ 个特征列；以及

- 30 第二训练模块 605，用于将包括所述 $(R+N-Q)$ 个特征列的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

在一优选实施例中，所述升维模块 603 用于：

利用 minwise Hashing 算法对所述 Q 个特征列进行升维处理，生成 R 个特征列，其中 $R=k*2^b$ ，k 和 b 为算法指定参数。

在一优选实施例中，所述机器学习算法模型为分类算法模型或回归算法模型。

5 在一优选实施例中，所述系统还包括：

第二降维模块 602a，用于当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，将所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $M < N$ ，且 $P < M$ ；

所述第二合并模块 604 用于：

10 所述将升维处理后生成的 R 个特征列和所述降维处理后生成的 P 个特征列合并。

根据本申请第六实施例提出的利用训练数据训练模型的系统，通过判断每一个特征列的重要程度值，采用将重要特征升维的方式，对于较小规模的训练数据，能够升高重要程度值较高的训练数据的维度，提高模型训练的准确性的问题。

15 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。

20

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号存储。信号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

25

30

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号。按照本文中的界定, 计算机可读介质不包括非持续性的电
5 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
10 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
15 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
20 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
25 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
30 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。

- 5 以上对本申请所提供的一种利用训练数据训练模型的方法和系统，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. 一种利用训练数据训练模型的方法，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，其特征不在于，所述训练方法包括：

5 计算每一个特征列的重要程度值；

判断每一个特征列的重要程度值是否小于对应的阈值；

当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，将所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $M < N$ ，且 $P < M$ ；

10 将重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列和降维处理后生成的 P 个特征列合并，得到 $(N-M+P)$ 个特征列；以及

将包括所述 $(N-M+P)$ 个特征列的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

2. 如权利要求 1 所述的训练模型的方法，其特征不在于，当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，将所述多个特征列进行降维处理，生成 P 个特征列的步骤包括：

15 当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，利用 minwise Hashing 算法对所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $P=k*2^b$ ， k 和 b 为算法指定参数。

3. 如权利要求 1 所述的训练模型的方法，其特征不在于，所述训练数据还包括对应于每一个样本的标签值，所述计算每一个特征列的重要程度值的步骤包括：

20 利用每一个样本对应的标签值和所述特征列中的多个特征，计算出所述特征列的信息值和/或信息增益，以所述信息值和/或信息增益作为重要程度值。

4. 如权利要求 3 所述的训练模型的方法，其特征不在于，所述信息值对应的阈值为第一阈值，所述信息增益对应的阈值为第二阈值。

25 5. 如权利要求 1 所述的训练模型的方法，其特征不在于，所述机器学习算法模型为分类算法模型或回归算法模型。

6. 一种利用训练数据训练模型的方法，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，其特征不在于，所述训练方法包括：

30 计算每一个特征列的重要程度值；

判断每一个特征列的重要程度值是否大于对应的阈值；

当判断出所述 N 个特征列中的 Q 个特征列的重要程度值大于对应的阈值时，将所述 Q 个特征列进行升维处理，生成 R 个特征列，其中 $R > Q$ ，且 $Q < N$ ；

将升维处理后生成的 R 个特征列和重要程度值小于或等于对应的阈值的特征列合并；以及

将合并的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

7. 如权利要求 6 所述的训练模型的方法，其特征在于，当判断出所述 N 个特征列中的 Q 个特征列的重要程度值大于对应的阈值时，将所述 Q 个特征列进行升维处理，生成 R 个特征列的步骤包括：

10 当判断出所述 N 个特征列中的 Q 个特征列的重要程度值大于对应的阈值时，利用 minwise Hashing 算法对所述 Q 个特征列进行升维处理，生成 R 个特征列，其中 $R = k * 2^b$ ， k 和 b 为算法指定参数。

8. 如权利要求 6 所述的训练模型的方法，其特征在于，所述训练数据还包括对应于每一个样本的标签值，所述计算每一个特征列的重要程度值的步骤包括：

15 利用每一个样本对应的标签值和所述特征列中的多个特征，计算出所述特征列的信息值和/或信息增益，以所述信息值和/或信息增益作为重要程度值。

9. 如权利要求 6 所述的训练模型的方法，其特征在于，所述判断每一个特征列的重要程度值是否大于对应的阈值的步骤之后，所述方法还包括：

20 当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，将所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $M < N$ ，且 $P < M$ ；

所述将升维处理后生成的 R 个特征列和重要程度值小于或等于对应的阈值的特征列合并的步骤包括：

将升维处理后生成的 R 个特征列和所述降维处理后生成的 P 个特征列合并。

25 10. 一种利用训练数据训练模型的系统，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，其特征在于，所述训练系统包括：

第一计算模块，用于计算每一个特征列的重要程度值；

第一判断模块，用于判断每一个特征列的重要程度值是否小于对应的阈值；

30 第一降维模块，用于当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对应的阈值时，将所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $M < N$ ，且 $P < M$ ；

第一合并模块，用于将重要程度值大于或等于对应的阈值的 $(N-M)$ 个特征列和降维处理后生成的 P 个特征列合并，得到 $(N-M+P)$ 个特征列；以及

第一训练模块，用于将包括所述 $(N-M+P)$ 个特征列的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

5 11. 如权利要求 10 所述的训练模型的系统，其特征在于，所述降维模块用于：

利用 minwise Hashing 算法对所述 M 个特征列进行降维处理，生成 P 个特征列，其中 $P=k*2^b$ ， k 和 b 为算法指定参数。

12. 如权利要求 10 所述的训练模型的系统，其特征在于，所述训练数据还包括对应于每一个样本的标签值，所述第一计算模块包括：

10 信息值计算模块，用于利用每一个样本对应的标签值和所述特征列中的多个特征，计算出所述特征列的信息值，以所述信息值作为所述重要程度值；和/或

信息增益计算模块，用于利用每一个样本对应的标签值和所述特征列中的多个特征，计算出所述特征列的信息增益，以所述信息增益作为所述重要程度值。

15 13. 如权利要求 12 所述的训练模型的系统，其特征在于，所述信息值对应的阈值为第一阈值，所述信息增益对应的阈值为第二阈值。

14. 如权利要求 10 所述的训练模型的系统，其特征在于，所述机器学习算法模型为分类算法模型或回归算法模型。

20 15. 一种利用训练数据训练模型的系统，所述训练数据包括多个样本，每一个样本包括 N 个特征，所述多个样本中对应的特征构成 N 个特征列，其特征在于，所述训练系统包括：

第二计算模块，用于计算每一个特征列的重要程度值；

第二判断模块，用于判断每一个特征列的重要程度值是否大于对应的阈值；

升维模块，用于当判断出所述 N 个特征列中的 Q 个特征列的重要程度值大于对应的阈值时，将所述 Q 个特征列进行升维处理，生成 R 个特征列，其中 $R>Q$ ，且 $Q<N$ ；

25 第二合并模块，用于将升维处理后生成的 R 个特征列和重要程度值小于或等于对应的阈值的 $(N-Q)$ 个特征列合并，得到 $(R+N-Q)$ 个特征列；以及

第二训练模块，用于将包括所述 $(R+N-Q)$ 个特征列的训练数据输入机器学习算法模型，训练所述机器学习算法模型。

16. 如权利要求 15 所述的训练模型的系统，其特征在于，所述升维模块用于：

30 利用 minwise Hashing 算法对所述 Q 个特征列进行升维处理，生成 R 个特征列，其

中 $R=k*2^b$, k 和 b 为算法指定参数。

17. 如权利要求 15 所述的训练模型的系统, 其特征在于, 所述训练数据还包括对应于每一个样本的标签值, 所述计算模块包括:

信息值计算模块, 用于利用每一个样本对应的标签值和所述特征列中的多个特征,

5 计算出所述特征列的信息值, 以所述信息值作为所述重要程度值; 和/或

信息增益计算模块, 用于利用每一个样本对应的标签值和所述特征列中的多个特征, 计算出所述特征列的信息增益, 以所述信息增益作为所述重要程度值。

18. 如权利要求 15 所述的训练模型的系统, 其特征在于, 所述系统还包括:

第二降维模块, 用于当判断出所述 N 个特征列中的 M 个特征列的重要程度值小于对
10 应的阈值时, 将所述 M 个特征列进行降维处理, 生成 P 个特征列, 其中 $M < N$, 且 $P < M$;

所述第二合并模块用于:

所述将升维处理后生成的 R 个特征列和所述降维处理后生成的 P 个特征列合并。

	特征列 1	特征列 2	特征列 3		特征列 N
样本 1	Value 1, 1	Value 1, 2	Value 1, 3		Value 1, n
样本 2	Value 2, 1	Value 2, 2	Value 2, 3		Value 2, n
样本 3	Value 3, 1	Value 3, 2	Value 3, 3		Value 3, n
样本 4	Value 4, 1	Value 4, 2	Value 4, 3		Value 4, n
样本 5	Value 5, 1	Value 5, 2	Value 5, 3		Value 5, n
.....					
样本 X	Value x, 1	Value x, 2	Value x, 3		Value x, n

图 1

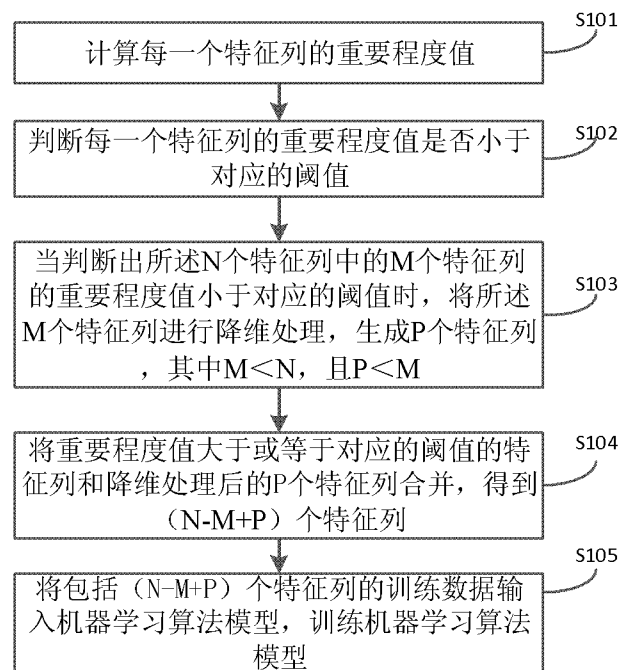


图2

col	IV	IG	col	IV	IG
1	5.87711....	2.20613...	100022	1.07377...	5.08820...
10	3.31695...	1.38631...	100024	6.04910...	2.46602...
1000	7.82036...	2.62107...	100026	1.67693...	5.96912...
100000	6.81132...	2.79715...	100028	6.99528...	3.59951...
100002	7.59792...	2.44489...	100031	1.76507...	8.72267...
100004	1.50500...	5.84963...	100033	5.96822...	2.53824...
100006	4.67386...	2.05491...	100035	2.58702...	9.15622...
100008	3.97857...	1.43503...	100037	1.62312...	7.11376...
100011	1.42644...	5.61722...	100039	6.58352...	2.74432...
100013	3.77851...	1.60399...	100040	1.32737...	5.41798...
100015	1.33614...	5.74715...	100042	5.18543...	2.22937...
100017	8.06321...	3.16101...	100044	3.23597...	1.50824...
100019	4.88318...	1.63962...	100046	1.81776...	8.01892...
100020	1.03414...	4.33213...			

图 3

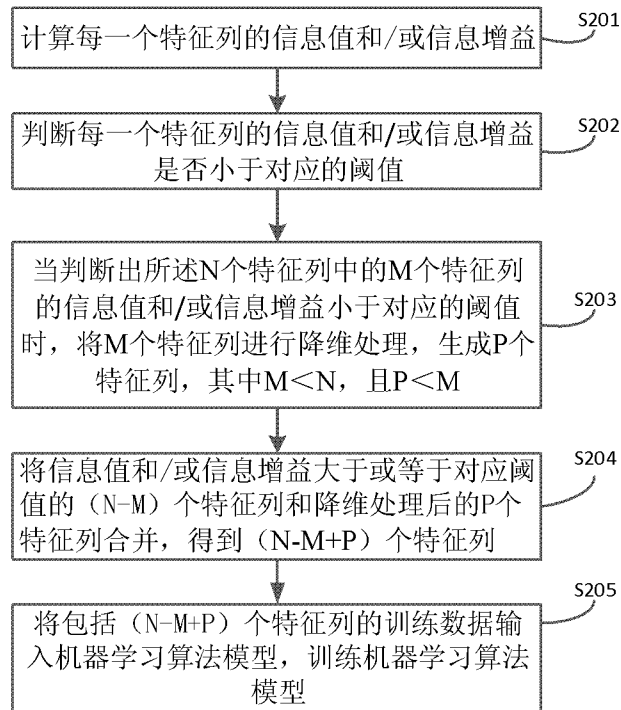


图4

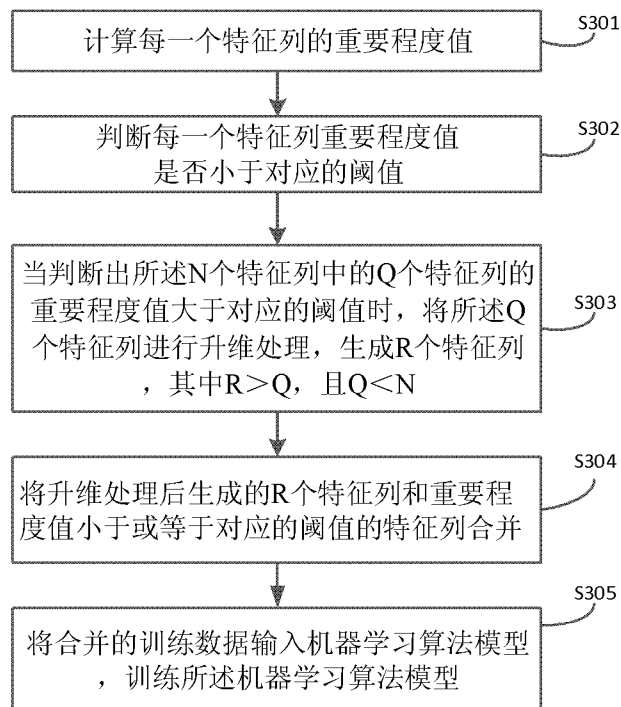


图5

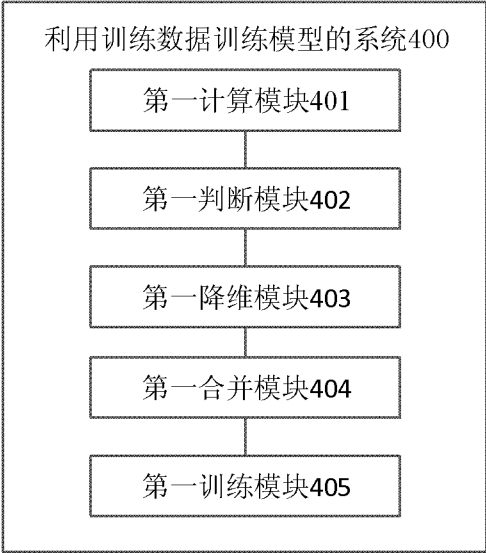


图6

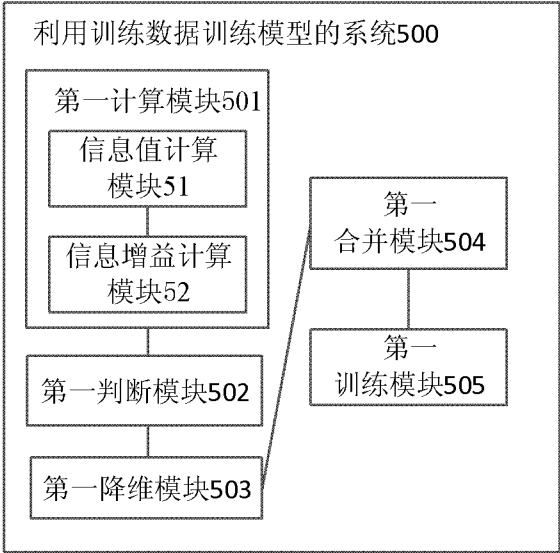


图7

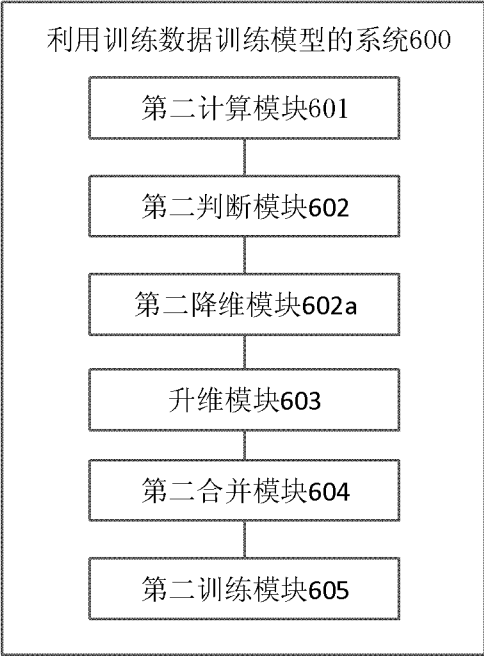


图8

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/073340

A. CLASSIFICATION OF SUBJECT MATTER

G06F 15/18 (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)

CNPAT, CNKI, WPI, EPODOC, IEEE: increase+, add+, model, dimension+, threshold, reduc+, train+, minhash, dimensionality
reduction, raising dimension

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 104050242 A (HARBIN UNIVERSITY OF SCIENCE AND TECHNOLOGY) 17 September 2014 (17.09.2014) description, paragraphs [0031]-[0102]	1-5, 10-14
Y	CN 104715021 A (NANJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS) 17 June 2015 (17.06.2015) description, paragraphs [0012]-[0023]	1-5, 10-14
A	CN 104239485 A (COMPUTER NETWORK INFORMATION CENTER, CHINESE ACADEMY OF SCIENCE) 24 December 2014 (24.12.2014) the whole document	1-18
A	CN 102609714 A (HARBIN UNIVERSITY OF SCIENCE AND TECHNOLOGY) 25 July 2012 (25.07.2012) the whole document	1-18
A	WO 2007096849 A1 (UNIVERSITY COLLEGE CORK-NATIONAL UNIVERSITY OF IRELAND, CORK) 30 August 2007 (30.08.2007) the whole document	1-18

☐ 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 13 April 2017	Date of mailing of the international search report 27 April 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 ZHANG, Qian Telephone No. (86-10) 62413681

International application No.
PCT/CN2017/073340

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104050242 A	17 September 2014	None	
CN 104715021 A	17 June 2015	None	
CN 104239485 A	24 December 2014	WO 2016033907 A1	10 March 2016
CN 102609714 A	25 July 2012	None	
WO 2007096849 A1	30 August 2007	IE 20070104 A1	03 October 2007

A. 主题的分类 G06F 15/18(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)) CNPAT, CNKI, WPI, EPDOC, IEEE: 增加, 提高, 模型, 升维, 维度, 阈值, 训练, 降维, increas+, add+, model, dimension+, threshold, reduc+, train+, minhash																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>Y</td> <td>CN 104050242 A (哈尔滨理工大学) 2014年 9月 17日 (2014 - 09 - 17) 说明书第[0031]-[0102]段</td> <td>1-5, 10-14</td> </tr> <tr> <td>Y</td> <td>CN 104715021 A (南京邮电大学) 2015年 6月 17日 (2015 - 06 - 17) 说明书第[0012]-[0023]段</td> <td>1-5, 10-14</td> </tr> <tr> <td>A</td> <td>CN 104239485 A (中国科学院计算机网络信息中心) 2014年 12月 24日 (2014 - 12 - 24) 全文</td> <td>1-18</td> </tr> <tr> <td>A</td> <td>CN 102609714 A (哈尔滨理工大学) 2012年 7月 25日 (2012 - 07 - 25) 全文</td> <td>1-18</td> </tr> <tr> <td>A</td> <td>WO 2007096849 A1 (UNIVERSITY COLLEGE CORK - NATIONAL UNIVERSITY OF IRELAND, CORK) 2007年 8月 30日 (2007 - 08 - 30) 全文</td> <td>1-18</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	Y	CN 104050242 A (哈尔滨理工大学) 2014年 9月 17日 (2014 - 09 - 17) 说明书第[0031]-[0102]段	1-5, 10-14	Y	CN 104715021 A (南京邮电大学) 2015年 6月 17日 (2015 - 06 - 17) 说明书第[0012]-[0023]段	1-5, 10-14	A	CN 104239485 A (中国科学院计算机网络信息中心) 2014年 12月 24日 (2014 - 12 - 24) 全文	1-18	A	CN 102609714 A (哈尔滨理工大学) 2012年 7月 25日 (2012 - 07 - 25) 全文	1-18	A	WO 2007096849 A1 (UNIVERSITY COLLEGE CORK - NATIONAL UNIVERSITY OF IRELAND, CORK) 2007年 8月 30日 (2007 - 08 - 30) 全文	1-18
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
Y	CN 104050242 A (哈尔滨理工大学) 2014年 9月 17日 (2014 - 09 - 17) 说明书第[0031]-[0102]段	1-5, 10-14																		
Y	CN 104715021 A (南京邮电大学) 2015年 6月 17日 (2015 - 06 - 17) 说明书第[0012]-[0023]段	1-5, 10-14																		
A	CN 104239485 A (中国科学院计算机网络信息中心) 2014年 12月 24日 (2014 - 12 - 24) 全文	1-18																		
A	CN 102609714 A (哈尔滨理工大学) 2012年 7月 25日 (2012 - 07 - 25) 全文	1-18																		
A	WO 2007096849 A1 (UNIVERSITY COLLEGE CORK - NATIONAL UNIVERSITY OF IRELAND, CORK) 2007年 8月 30日 (2007 - 08 - 30) 全文	1-18																		
☐ 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 13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 张千 电话号码 (86-10)62413681																		

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/073340

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	104050242	A	2014年 9月 17日	无			
CN	104715021	A	2015年 6月 17日	无			
CN	104239485	A	2014年 12月 24日	WO	2016033907	A1	2016年 3月 10日
CN	102609714	A	2012年 7月 25日	无			
WO	2007096849	A1	2007年 8月 30日	IE	20070104	A1	2007年 10月 3日