

(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 2017/143919 A1

- (51) 国际专利分类号:
G06K 9/62 (2006.01) G06Q 30/06 (2012.01)
G06N 3/08 (2006.01)
- (21) 国际申请号: PCT/CN2017/073444
- (22) 国际申请日: 2017 年 2 月 14 日 (14.02.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (30) 优先权:
201610110817.3 2016 年 2 月 26 日 (26.02.2016) 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: 姜晓燕 (JIANG, 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)。 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)。 褚威 (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)。

(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,

[见续页]

(54) Title: METHOD AND APPARATUS FOR ESTABLISHING DATA IDENTIFICATION MODEL

(54) 发明名称: 一种建立数据识别模型的方法及装置

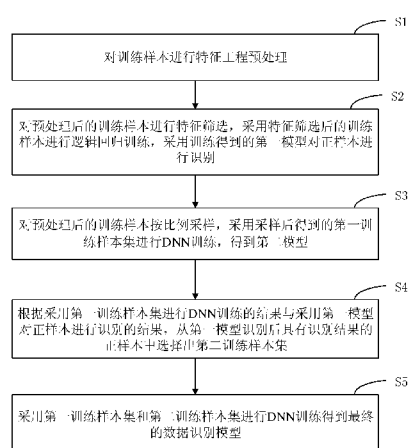


图 1

- S1 Perform feature engineering preprocessing on training samples
- S2 Perform feature screening on the preprocessed training samples, perform logistic regression training by using the training samples subjected to the feature screening, and identify positive samples using a first model obtained by the training
- S3 Proportionally sample the preprocessed training samples, and perform DNN training by using a first training sample set obtained by the sampling, to obtain a second model
- S4 Select, according to the result of performing DNN training by using the first training sample set and the result of identifying the positive samples using the first model, positive samples having identification results after being identified using the first model, to obtain a second training sample set
- S5 Perform DNN training by using the first training sample set and the second training sample set to obtain a final data identification model

(57) Abstract: A method and apparatus for establishing a data identification model, for use in establishing a data identification model according to training samples comprising positive and negative samples. The method comprises: performing feature engineering preprocessing on training samples (S1); performing feature screening on the preprocessed training samples, performing logistic regression training by using the training samples subjected to the feature screening, and identifying positive samples using a first model obtained by the training (S2); proportionally sampling the preprocessed training samples, and performing DNN training by using a first training sample set obtained by the sampling, to obtain a second model (S3); selecting, according to the result of performing DNN training by using the first training sample set and the result of identifying the positive samples using the first model, positive samples having identification results after being identified using the first model, to obtain a second training sample set (S4); and performing DNN training by using the first training sample set and the second training sample set to obtain a final data identification model (S5). The apparatus comprises a first training module, a sampling module, a selection module, and a final model training module. The present invention improves the stability of data identification models.

(57) 摘要:

[见续页]

WO 2017/143919 A1



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。

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))。

一种建立数据识别模型的方法及装置, 用于根据包括正、负样本的训练样本建立数据识别模型, 该方法包括: 对训练样本进行特征工程预处理 (S1); 对预处理后的训练样本进行特征筛选, 采用特征筛选后的训练样本进行逻辑回归训练, 采用训练得到的第一模型对正样本进行识别 (S2); 对预处理后的训练样本按比例采样, 采用采样后得到的第一训练样本集进行 DNN 训练, 得到第二模型 (S3); 根据采用第一训练样本集进行 DNN 训练的结果与采用第一模型对正样本进行识别的结果, 从第一模型识别后具有识别结果的正样本中选出第二训练样本集 (S4); 采用第一训练样本集和第二训练样本集进行 DNN 训练得到最终的数据识别模型 (S5)。装置包括第一训练模块、采样模块、选择模块和最终模型训练模块。提高了数据识别模型的稳定性。

一种建立数据识别模型的方法及装置

本申请要求 2016 年 02 月 26 日递交的申请号为 201610110817.3、发明名称为“一种建立数据识别模型的方法及装置”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

本发明属于数据处理技术领域，尤其涉及一种建立数据识别模型的方法及装置。

背景技术

10 商家的信用是消费者决定是否消费的重要指标，目前网上电商平台也是按照商家的信用高低进行排名。商家的信用根据交易的数量和评分逐步累积，刚开的店铺没有信用，排名就会靠后。消费者出于对自身权益的考虑，更愿意选择信用较高的商家或者销量较高的商品。而商家排名的先后直接关系到消费者是否能够搜索到商家，搜索不到的情况下，消费者就无法进入商家的店铺进行消费。

15 因此网上商家都有提升信用的需求，催生了一些专为商家提升信用的网站和个人，通过刷单等虚假交易行为来提升商家的信用。虚假交易行为不利于市场的健康发展，不利于保护消费者的权益，属于电商平台需要严厉打击的行为。

电商平台例如小微金服花呗和信贷业务，在使用时都要利用训练得到的识别模型来识别交易是否是虚假交易。通常在业务上通过 TOP 抓坏率来衡量对虚假交易的识别是否
20 准确，所谓抓坏率也称为召回率，是指识别出的虚假交易占虚假交易总数的比率。TOP 抓坏率是用于对训练得到的模型进行评估的指标，按模型识别得到的虚假交易概率对交易记录进行排序，然后对排序后的交易记录进行分组，计算各组的抓坏率，如果 TOP 抓坏率保持稳定且能达到设定的标准，则判断模型可靠，可用于后续的识别。

然而目前小微金服等电商平台在训练识别模型时，一般是先对训练样本通过特征工
25 程处理后，经过逻辑回归算法训练得到识别模型，然后采用测试样本来计算抓坏率，根据抓坏率来判断训练得到的识别模型是否可靠。

但是现在训练得到的识别模型是使用逻辑回归模型，对于训练样本按比例采样，没有对正样本进行区分，导致噪音进入逻辑回归算法，无法有效提高 TOP 抓坏率和保证稳定性。并且随着虚假交易维度越来越多，线性模型已经无法学到更多维度的信息，模型
30 单一，效果受限。

发明内容

本发明的目的是提供一种建立数据识别模型的方法及装置，以解决现有技术逻辑回归模型训练时噪音的影响，以及模型单一、效果不理想等问题。结合机器学习和深度学习进行训练，在判断虚假交易时，有效提高 TOP 抓坏率，取得很好的效果。

为了实现上述目的，本发明技术方案如下：

一种建立数据识别模型的方法，用于根据包括正、负样本的训练样本建立数据识别模型，所述建立数据识别模型的方法包括：

采用训练样本进行逻辑回归训练，得到第一模型；

10 对训练样本按比例采样，获得第一训练样本集；

采用训练得到的第一模型对正样本进行识别，从第一模型识别后具有识别结果的正样本中选择出第二训练样本集；

采用采样后得到的第一训练样本集与所述第二训练样本集进行深度神经网络 DNN 训练，得到最终的数据识别模型。

15 进一步地，所述建立数据识别模型的方法，在进行按比例采样或进行逻辑回归训练前，还包括：

对训练样本进行特征工程预处理。

进一步地，所述建立数据识别模型的方法，在采用训练样本进行逻辑回归训练之前，还包括：

20 对训练样本进行特征筛选，所述特征筛选通过计算特征的信息值，去除信息值小于设定阈值的特征。

优选地，所述从第一模型识别后具有识别结果的正样本中选择出第二训练样本集之前，还包括：

采用第一训练样本集进行 DNN 训练，得到第二模型。

25 进一步地，所述从第一模型识别后具有识别结果的正样本中选择出第二训练样本集，包括：

对训练得到的第一模型进行评估，得到第一模型对应的 ROC 曲线；

对训练得到的第二模型进行评估，得到第二模型对应的 ROC 曲线；

30 根据第一模型与第二模型 ROC 曲线的交点对应的阈值概率，从第一模型识别后具有识别结果的正样本中选择出概率小于所述阈值概率的样本作为第二训练样本集。

本发明优选地选择第二训练样本集的方法能够选择出更加符合训练要求的样本，提高最终数据识别模型的稳定性。

本发明还提出了一种建立数据识别模型的装置，用于根据包括正、负样本的训练样本建立数据识别模型，所述装置包括：

- 5 第一训练模块，用于采用训练样本进行逻辑回归训练，得到第一模型；
 采样模块，用于对训练样本按比例采样，获得第一训练样本集；
 选择模块，用于采用训练得到的第一模型对正样本进行识别，从第一模型识别后具有识别结果的正样本中选择出第二训练样本集；

- 最终模型训练模块，用于采用采样后得到的第一训练样本集与所述第二训练样本集
10 进行深度神经网络 DNN 训练，得到最终的数据识别模型。

进一步地，所述装置还包括：

 预处理模块，用于在进行按比例采样或进行逻辑回归训练前，对训练样本进行特征工程预处理。

进一步地，所述装置还包括：

- 15 特征筛选模块，用于在采用训练样本进行逻辑回归训练之前，对训练样本进行特征筛选，所述特征筛选通过计算特征的信息值，去除信息值小于设定阈值的特征。

优选地，本发明所述装置还包括：

 第二训练模块，用于采用第一训练样本集进行 DNN 训练，得到第二模型。

- 进一步地，所述选择模块从第一模型识别后具有识别结果的正样本中选择出第二训练样本集时，执行如下操作：
20

 对训练得到的第一模型进行评估，得到第一模型对应的 ROC 曲线；

 对训练得到的第二模型进行评估，得到第二模型对应的 ROC 曲线；

 根据第一模型与第二模型 ROC 曲线的交点对应的阈值概率，从第一模型识别后具有识别结果的正样本中选择出概率小于所述阈值概率的样本作为第二训练样本集。

- 25 本发明提出的一种建立数据识别模型的方法及装置，通过对全部训练样本进行特征工程预处理以及特征筛选，并根据逻辑回归训练得到的第一模型识别结果和采用第一训练样本集进行 DNN 训练的结果，从具有识别结果的所有正样本中选择出第二训练样本集，来结合深度神经网络训练得到最终的数据识别模型，提高了模型的稳定性。

图 1 为本发明建立数据识别模型的方法流程图；

图 2 为本发明数据识别模型评估效果对照图；

图 3 为本发明建立数据识别模型的装置结构示意图。

5 具体实施方式

下面结合附图和实施例对本发明技术方案做进一步详细说明，以下实施例不构成对本发明的限定。

如图 1 所示，本实施例一种建立数据识别模型的方法，包括：

步骤 S1、对训练样本进行特征工程预处理。

10 对于获取的全部训练样本，由于样本中的特征有些值缺失，或者偏差超出正常的范围，会影响到后续的训练，通常需要对样本进行特征工程处理。本实施例首先对样本进行特征工程预处理，即对样本的特征进行数据替换和清洗，剔除无意义特征。例如对样本中缺失的特征进行数据替换等。

15 步骤 S2、对预处理后的训练样本进行特征筛选，采用特征筛选后的训练样本进行逻辑回归训练，采用训练得到的第一模型对正样本进行识别。

全部训练样本中包括正样本和负样本，本实施例以虚假交易为例来进行说明，正样本表示是虚假交易的样本，负样本表示不是虚假交易的样本。

20 在模型识别中，因为有些特征与最终识别结果关系不大，若把这些特征作为变量会使得模型识别结果变差，或一般情况下应使特征数大大小于样本数，所以有必要采用特征筛选来筛选掉不重要甚至有负作用的特征。进行特征筛选的方法很多，例如有最近邻算法、偏最小二乘法等。本实施例优选地通过采用信息值 IV (information value) 来对样本的特征进行筛选。通过计算样本每个特征对应的信息值，将特征对应的信息值小于设定阈值的样本特征去除，减少其对样本分布的影响。

25 本实施例计算样本特征对应的信息值是根据所有训练样本的特征来计算，假设一条训练样本的特征包括 {feature 1、feature 2、...、feature m}，对于其中的一个特征 feature i，i 属于 $(1 \sim m)$ ，m 为特征数量。所有训练样本对应该 feature i 的值为 {i1, i2, ..., in}，n 为训练样本总数。

则可以根据 feature i 的值进行分组，例如将 feature i 的值为 a 的划分为一组，这样将 feature i 分为 K 组，根据如下公式计算特征 feature i 的信息值 IV：

$$WOE_{ki} = \log\left(\frac{Disgood_{ki}}{Disbad_{ki}}\right)$$

$$IV = \sum_{k=1}^K (Disgood_{ki} - Disbad_{ki}) * WOE_{ki}$$

其中，Disgood_{ki} 为样本组中负样本数量，Disbad_{ki} 为样本组中正样本数量。本实施例不限定哪个为负样本数量，哪个为正样本数量，即也可以用 Disgood_{ki} 表示正样本数量，Disbad_{ki} 表示负样本数量。从而可以根据特征对应的信息值来筛选特征，将对应信息值小于设定阈值的特征舍弃，保留对结果有影响的特征用来进行后续的训练，提高训练模型的可靠性。

在进行特征筛选后，采用特征筛选后的全部训练样本进行逻辑回归训练得到第一模型，该模型即为现有技术方案中采用的识别模型。本发明在此基础上进一步训练以得到更加可靠的模型。一般来说采用特征筛选后的全部训练样本进行逻辑回归训练得到第一模型稳定性比较好，可以选择其中的一些样本来进行后续的训练，以使得后续训练得到的模型具有较好的稳定性。衡量模型稳定性一般采用 TOP 抓坏率指标，TOP 抓坏率可以根据模型识别样本得到的虚假交易概率来进行计算。

为此，本实施例采用训练得到的第一模型对所有正样本进行识别，得到每个训练样本对应的为虚假交易的概率，记所有正样本及其识别得到的概率为训练集合 B，即通过第一模型识别后具有识别结果的正样本。在后续步骤中根据识别结果从训练集合 B 中选择一部分训练样本作为后续的训练用。

步骤 S3、对预处理后的训练样本按比例采样，采用采样后得到的第一训练样本集进行 DNN 训练，得到第二模型。

为了从训练集合 B 中选择一部分训练样本作为后续的训练用，可以直接从训练集合 B 中选择识别准确的样本作为后续训练采用的第二训练样本集。

本实施例优选地对预处理后的全部训练样本按比例采样得到训练集合 A（第一训练样本集），例如正负样本的比例为 1: 10。在操作中，先选择出所有的正样本，然后从负样本中选择足够多的负样本，保持 1: 10 的比例。然后采用采样后得到的第一训练样本集进行 DNN 训练，可以得到一个第二模型。深度神经网络 DNN (Deep Neural Networks) 是近年来机器学习领域中的研究热点，DNN 训练广泛应用于语音识别及其他数据分类上，关于 DNN 训练的内容这里不再赘述。

在后续步骤中根据第二模型的训练结果与第一模型的训练结果从训练集合 B 中选择

第二训练样本集。

根据实验得到的经验，第二模型的识别结果稳定性不够。而结合第二训练样本集在后续步骤中进行训练能够得到稳定性好的最终数据识别模型。

需要说明的是，本实施例对全部训练样本进行特征工程预处理，以及采用特征筛选来筛选掉不重要甚至有负作用的特征，都是为了训练得到的模型更加可靠。在具体的实施例中，可以在训练得到第一模型和训练得到第二模型时都需要对训练样本进行预处理和特征筛选，也可以仅在训练得到第一模型时进行特征筛选，而在训练第二模型时不进行特征筛选。容易理解的是，即使不进行特征工程预处理及特征筛选，也能提高训练得到的模型的识别效果，使得训练得到的模型的识别效果好于现有技术，这里不再赘述。

步骤 S4、根据采用第一训练样本集进行 DNN 训练的结果与采用第一模型对正样本进行识别的结果，从第一模型识别后具有识别结果的正样本中选择出第二训练样本集。

ROC 曲线是显示模型真正率和假正率的一种图形化方法，常用来评估模型的效果，ROC 曲线上每个点对应有三个值，分别为纵坐标真正率 (True Positive Rate, TPR)、横坐标假正率 (False Positive Rate, FPR) 和阈值概率。真正率 (True Positive Rate, TPR) 是指被模型预测为正的正样本与正样本实际数量的比率；假正率 (False Positive Rate, FPR) 是指被模型预测为正的负样本与负样本实际数量的比率；阈值概率是用来判定预测结果为正的判定阈值，如果样本预测的结果大于该阈值概率则判定为正，否则判定为负。模型的预测效果越好，其 TPR 越接近于 1，FPR 越接近于 0。

本实施例从训练集合 B 中选择一部分训练样本作为后续的训练用，选择的具体方法包括：

对训练得到的第二模型进行评估，得到第二模型对应的 ROC 曲线；

对训练得到的第一模型进行评估，得到第一模型对应的 ROC 曲线；

根据第一模型与第二模型 ROC 曲线的交点对应的阈值概率，选择训练集合 B 中概率小于该阈值概率的样本，作为第二训练样本集。

需要说明的是，选择的第二训练样本集中的样本数量小于第一训练样本集中的正样本数量，最多不超过第一训练样本集中的正样本数量，这样是为了保证正负样本的比例，以防止正样本过多导致模型整体效果变差。

选择第二训练样本集还可以根据模型评估得到的概率，从训练集合 B 中按照概率从大到小顺序选择一定数量的样本第二训练样本集。或者根据经验设定一个阈值，从训练集合 B 中选择概率大于该阈值的样本作为第二训练样本集。本发明优选地根据 ROC 曲线

的交点进行选择，能够保证在后续的训练中得到更好的结果。

步骤 S5、采用第一训练样本集和第二训练样本集进行 DNN 训练得到最终的数据识别模型。

最后采用第一训练样本集和第二训练样本集进行 DNN 训练得到最终的数据识别模型，关于 DNN 深度学习训练模型，这里不再赘述。如图 2 所示的 ROC 曲线表明，本实施例训练得到的最终的数据识别模型效果远远好于直接通过逻辑回归训练得到的第一模型效果。图 2 中上面的曲线为本实施例训练得到的最终的数据识别模型对应的 ROC 曲线，下面的曲线为直接通过逻辑回归训练得到的第一模型对应的 ROC 曲线。

通过对最终数据识别模型 TOP 抓坏率的计算，可以发现本实施例提出的建立数据识别模型的方法大大提高了模型的稳定性。

如图 3 所示，本实施例还提出了一种建立数据识别模型的装置，用于根据包括正、负样本的训练样本建立数据识别模型，该装置包括：

第一训练模块，用于采用训练样本进行逻辑回归训练，得到第一模型；

采样模块，用于对训练样本按比例采样，获得第一训练样本集；

选择模块，用于采用训练得到的第一模型对正样本进行识别，从第一模型识别后具有识别结果的正样本中选择出第二训练样本集；

最终模型训练模块，用于采用采样后得到的第一训练样本集与所述第二训练样本集进行深度神经网络 DNN 训练，得到最终的数据识别模型。

与上述方法对应地，容易理解的是，本装置还包括：

预处理模块，用于在进行按比例采样或进行逻辑回归训练前，对训练样本进行特征工程预处理。

以及，本装置还包括：

特征筛选模块，用于在采用训练样本进行逻辑回归训练之前，对训练样本进行特征筛选，所述特征筛选通过计算特征的信息值，去除信息值小于设定阈值的特征。

优选地，本装置还包括：

第二训练模块，用于采用第一训练样本集进行 DNN 训练，得到第二模型。

则本实施例采用优选的方法来选择第二训练数据集，选择模块从第一模型识别后具有识别结果的正样本中选择出第二训练样本集时，执行如下操作：

对训练得到的第一模型进行评估，得到第一模型对应的 ROC 曲线；

对训练得到的第二模型进行评估，得到第二模型对应的 ROC 曲线；

根据第一模型与第二模型 ROC 曲线的交点对应的阈值概率，从第一模型识别后具有识别结果的正样本中选择出概率小于所述阈值概率的样本作为第二训练样本集。

以上实施例仅用以说明本发明的技术方案而非对其进行限制，在不背离本发明精神及其实质的情况下，熟悉本领域的技术人员当可根据本发明作出各种相应的改变和变形，但这些相应的改变和变形都应属于本发明所附的权利要求的保护范围。

权利要求书

1. 一种建立数据识别模型的方法，用于根据包括正、负样本的训练样本建立数据识别模型，其特征在于，所述建立数据识别模型的方法包括：

采用训练样本进行逻辑回归训练，得到第一模型；

5 对训练样本按比例采样，获得第一训练样本集；

采用训练得到的第一模型对正样本进行识别，从第一模型识别后具有识别结果的正样本中选择出第二训练样本集；

采用采样后得到的第一训练样本集与所述第二训练样本集进行深度神经网络 DNN 训练，得到最终的数据识别模型。

10 2. 根据权利要求 1 所述的建立数据识别模型的方法，其特征在于，所述建立数据识别模型的方法，在进行按比例采样或进行逻辑回归训练前，还包括：

对训练样本进行特征工程预处理。

3. 根据权利要求 2 所述的建立数据识别模型的方法，其特征在于，所述建立数据识别模型的方法，在采用训练样本进行逻辑回归训练之前，还包括：

15 对训练样本进行特征筛选，所述特征筛选通过计算特征的信息值，去除信息值小于设定阈值的特征。

4. 根据权利要求 1 所述的建立数据识别模型的方法，其特征在于，所述从第一模型识别后具有识别结果的正样本中选择出第二训练样本集之前，还包括：

采用第一训练样本集进行 DNN 训练，得到第二模型。

20 5. 根据权利要求 4 所述的建立数据识别模型的方法，其特征在于，所述从第一模型识别后具有识别结果的正样本中选择出第二训练样本集，包括：

对训练得到的第一模型进行评估，得到第一模型对应的 ROC 曲线；

对训练得到的第二模型进行评估，得到第二模型对应的 ROC 曲线；

25 根据第一模型与第二模型 ROC 曲线的交点对应的阈值概率，从第一模型识别后具有识别结果的正样本中选择出概率小于所述阈值概率的样本作为第二训练样本集。

6. 一种建立数据识别模型的装置，用于根据包括正、负样本的训练样本建立数据识别模型，其特征在于，所述装置包括：

第一训练模块，用于采用训练样本进行逻辑回归训练，得到第一模型；

采样模块，用于对训练样本按比例采样，获得第一训练样本集；

30 选择模块，用于采用训练得到的第一模型对正样本进行识别，从第一模型识别后具

有识别结果的正样本中选择出第二训练样本集；

最终模型训练模块，用于采用采样后得到的第一训练样本集与所述第二训练样本集进行深度神经网络 DNN 训练，得到最终的数据识别模型。

7. 根据权利要求 6 所述的建立数据识别模型的装置，其特征在于，所述装置还包括：

5 预处理模块，用于在进行按比例采样或进行逻辑回归训练前，对训练样本进行特征工程预处理。

8. 根据权利要求 7 所述的建立数据识别模型的装置，其特征在于，所述装置还包括：

特征筛选模块，用于在采用训练样本进行逻辑回归训练之前，对训练样本进行特征筛选，所述特征筛选通过计算特征的信息值，去除信息值小于设定阈值的特征。

10 9. 根据权利要求 6 所述的建立数据识别模型的装置，其特征在于，所述装置还包括：第二训练模块，用于采用第一训练样本集进行 DNN 训练，得到第二模型。

10. 根据权利要求 9 所述的建立数据识别模型的装置，其特征在于，所述选择模块从第一模型识别后具有识别结果的正样本中选择出第二训练样本集时，执行如下操作：

对训练得到的第一模型进行评估，得到第一模型对应的 ROC 曲线；

15 对训练得到的第二模型进行评估，得到第二模型对应的 ROC 曲线；

根据第一模型与第二模型 ROC 曲线的交点对应的阈值概率，从第一模型识别后具有识别结果的正样本中选择出概率小于所述阈值概率的样本作为第二训练样本集。

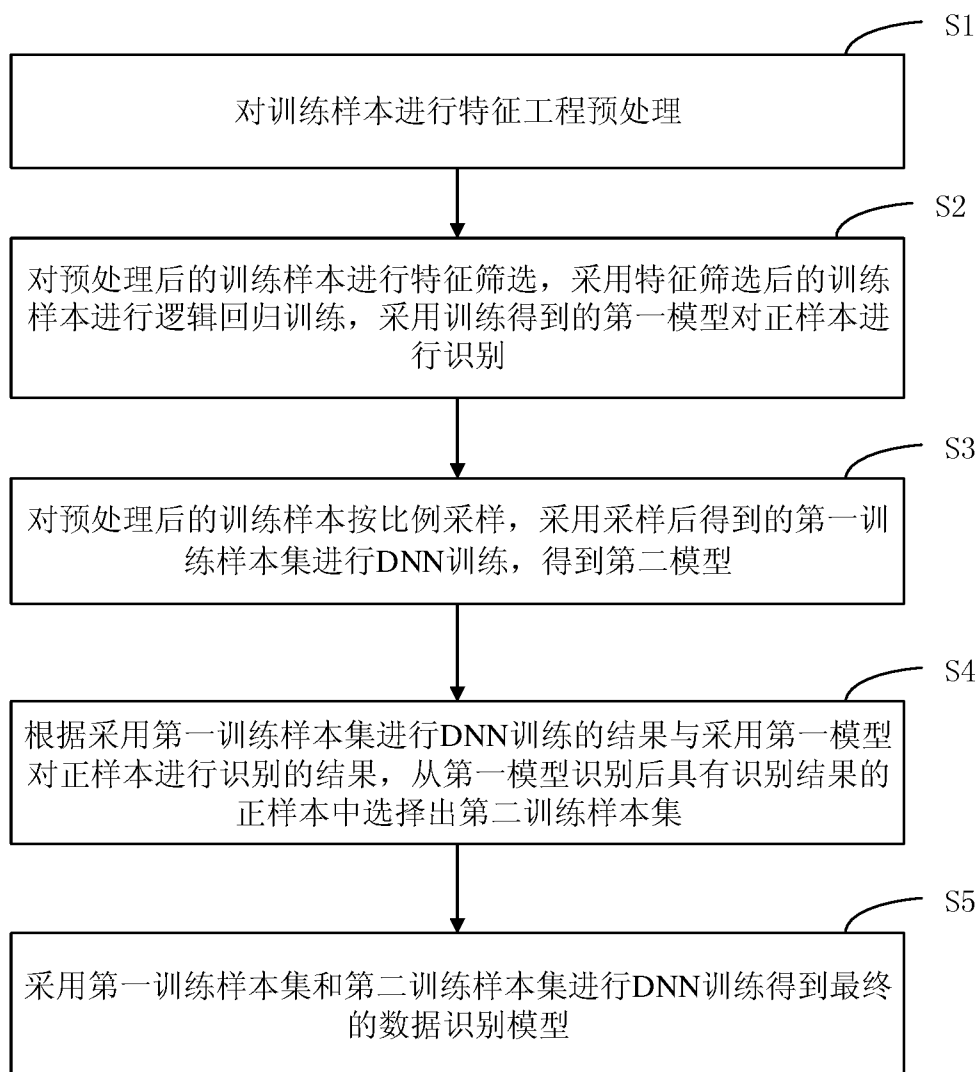


图 1

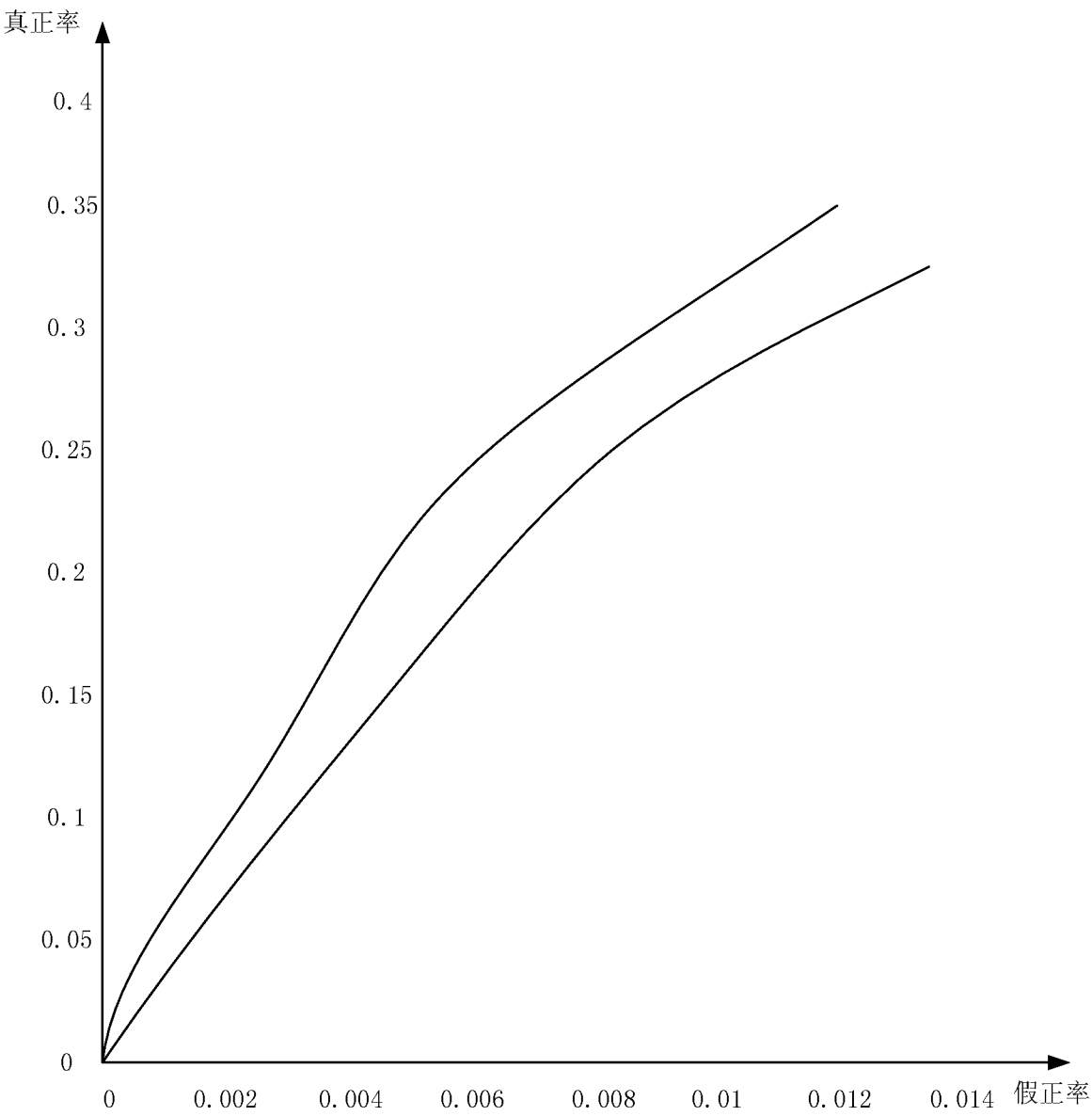


图 2

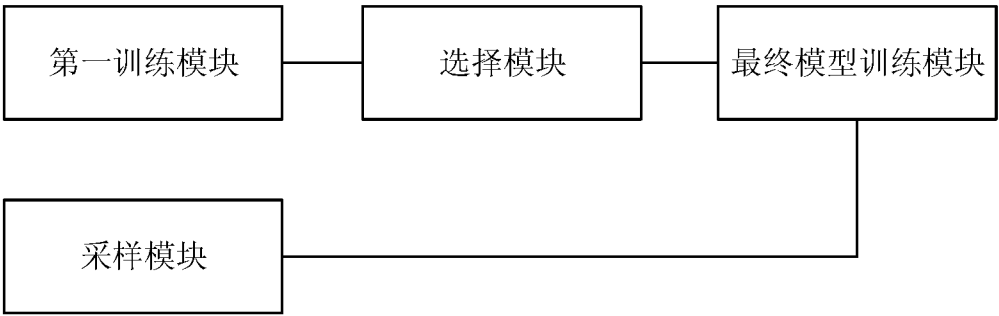


图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/073444

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62 (2006.01) i; G06N 3/08 (2006.01) i; G06Q 30/06 (2012.01) 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; G06N; G06Q

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, CNKI, CNPAT, GOOGLE, IEEE: negative, positive, sample, neural w network, model, training, DNN, data, recognition, sample, ROC, threshold, proportion, regression

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 104966097 A (CHENGDU BUSINESS BIG DATA TECHNOLOGY CO., LTD.) 07 October 2015 (07.10.2015) description, paragraphs [0066]-[0077]	1-4, 6-9
A	CN 103902968 A (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY) 02 July 2014 (02.07.2014) the whole document	1-10
A	CN 101799875 A (HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY) 11 August 2010 (11.08.2010) the whole document	1-10
A	US 2015095017 A1 (GOOGLE INC.) 02 April 2015 (02.04.2015) the whole document	1-10

☐ 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 06 April 2017	Date of mailing of the international search report 08 May 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, Tao Telephone No. (86-10) 62414427

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/073444

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 104966097 A	07 October 2015	None	
CN 103902968 A	02 July 2014	None	
CN 101799875 A	11 August 2010	None	
US 2015095017 A1	02 April 2015	None	

A. 主题的分类 G06K 9/62(2006.01)i; G06N 3/08(2006.01)i; G06Q 30/06(2012.01)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; G06N; G06Q 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, GOOGLE, IEEE: 正样本, 负样本, 神经网络, 模型, 训练, 数据, 识别, 采样, 阈值, 比例, 回归, negative, positive, sample, neural w network, model, training, DNN, data, recognition, sampling, ROC, threshold, proportion, regression		
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
X	CN 104966097 A (成都数联铭品科技有限公司) 2015年 10月 7日 (2015 - 10 - 07) 说明书第[0066]-[0077]段	1-4、6-9
A	CN 103902968 A (中国人民解放军国防科学技术大学) 2014年 7月 2日 (2014 - 07 - 02) 全文	1-10
A	CN 101799875 A (华中科技大学) 2010年 8月 11日 (2010 - 08 - 11) 全文	1-10
A	US 2015095017 A1 (GOOGLE INC.) 2015年 4月 2日 (2015 - 04 - 02) 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 2017年 4月 6日		国际检索报告邮寄日期 2017年 5月 8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 张涛 电话号码 (86-10)62414427

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/073444

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	104966097	A	2015年 10月 7日	无	
CN	103902968	A	2014年 7月 2日	无	
CN	101799875	A	2010年 8月 11日	无	
US	2015095017	A1	2015年 4月 2日	无	