



(12)发明专利申请

(10)申请公布号 CN 110232150 A

(43)申请公布日 2019. 09. 13

(21)申请号 201910423207.2

(22)申请日 2019.05.21

(71)申请人 平安科技(深圳)有限公司

地址 518000 广东省深圳市福田区福田街  
道福安社区益田路5033号平安金融中  
心23楼

(72)发明人 戴方慧 黄薇

(74)专利代理机构 深圳中一专利商标事务所

44237

代理人 冷仔

(51)Int.Cl.

G06F 16/9535(2019.01)

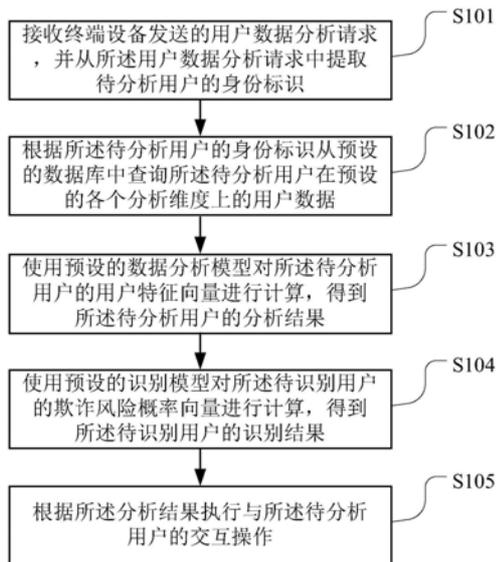
权利要求书4页 说明书11页 附图4页

(54)发明名称

一种用户数据分析方法、装置、可读存储介  
质及终端设备

(57)摘要

本发明属于计算机技术领域,尤其涉及一种  
用户数据分析方法、装置、计算机可读存储介质  
及终端设备。所述方法接收终端设备发送的用户  
数据分析请求,从中提取待分析用户的身份标  
识;根据所述待分析用户的身份标识从预设的数  
据库中查询所述待分析用户在预设的各个分析  
维度上的用户数据;根据各个分析维度上的用户  
数据构造所述待分析用户的用户特征向量;使用  
预设的数据分析模型对所述待分析用户的用户  
特征向量进行计算,得到所述待分析用户的分析  
结果,所述数据分析模型为经过预设的样本集合  
训练的机器学习模型;根据所述分析结果执行与  
所述待分析用户的交互操作。对用户数据进行充  
分的挖掘分析,以尽量满足用户的需求,从而极  
大提升了用户的体验。



1. 一种用户数据分析方法,其特征在于,包括:

接收终端设备发送的用户数据分析请求,并从所述用户数据分析请求中提取待分析用户的身份标识;

根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据;

根据各个分析维度上的用户数据构造所述待分析用户的用户特征向量;

使用预设的数据分析模型对所述待分析用户的用户特征向量进行计算,得到所述待分析用户的分析结果,所述数据分析模型为经过预设的样本集合训练的机器学习模型;

根据所述分析结果执行与所述待分析用户的交互操作。

2. 根据权利要求1所述的用用户数据分析方法,其特征在于,所述数据分析模型的构建过程包括:

根据下式构建所述数据分析模型:

$$\text{CalRes}_u = \text{Sigmoid}(\text{CoefVec} \times \text{UserVec}_u^T)$$

其中,Sigmoid为S型生长曲线函数, $u$ 为所述样本集合中的历史用户的序号, $1 \leq u \leq UN$ , $\text{UserVec}_u$ 为所述样本集合中第 $u$ 个历史用户的用户特征向量,且 $\text{UserVec}_u = (\text{UsData}_{u,1}, \text{UsData}_{u,2}, \dots, \text{UsData}_{u,d}, \dots, \text{UsData}_{u,DN})$ , $\text{UsData}_{u,d}$ 为所述样本集合中第 $u$ 个历史用户在第 $d$ 个分析维度上的用户数据, $\text{CoefVec}$ 为预设的第一模型系数向量,且 $\text{CoefVec} = (\text{Coef}_1, \text{Coef}_2, \dots, \text{Coef}_d, \dots, \text{Coef}_{DN})$ , $\text{Coef}_d$ 为第 $d$ 个分析维度上的权重系数, $T$ 为转置符号, $\text{CalRes}_u$ 为所述样本集合中第 $u$ 个历史用户的模型输出值;

使用所述样本集合对所述数据分析模型进行训练,求解出所述第一模型系数向量;

从所述第一模型系数向量中选取 $G$ 个权重系数,并构建如下所示的第二模型系数向量:

$$\text{NwCoefVec} = (\text{SqCoef}_1, \text{SqCoef}_2, \dots, \text{SqCoef}_{nd}, \dots, \text{SqCoef}_G)$$

其中, $\text{NwCoefVec}$ 为所述第二模型系数向量, $nd$ 为所述第二模型系数向量中的各个权重系数的序号, $1 \leq nd \leq G$ , $G$ 为正整数, $\text{SqCoef}_{nd}$ 为所述第二模型系数向量中的第 $nd$ 个权重系数;

根据下式对所述数据分析模型进行更新:

$$\text{CalRes}_u = \text{Sigmoid}(\text{NwCoefVec} \times \text{NwUserVec}_u^T)$$

其中, $\text{NwUserVec}_u = (\text{NwUsData}_{u,1}, \text{NwUsData}_{u,2}, \dots, \text{NwUsData}_{u,nd}, \dots, \text{NwUsData}_{u,G})$ , $\text{NwUsData}_{u,s}$ 为所述样本集合中第 $u$ 个历史用户与所述第二模型系数向量中的第 $nd$ 个权重系数对应的分析维度上的用户数据;

使用所述样本集合对更新后的所述数据分析模型进行训练,求解出所述第二模型系数向量。

3. 根据权利要求2所述的用用户数据分析方法,其特征在于,所述从所述第一模型系数向量中选取 $G$ 个权重系数包括:

将所述第一模型系数向量中的各个权重系数按照从大到小的顺序排列为如下所示的系数序列:

$$\text{SqCoef}_1, \text{SqCoef}_2, \dots, \text{SqCoef}_s, \dots, \text{SqCoef}_{DN}$$

其中, $s$ 为所述第一模型系数向量中的各个权重系数按照从大到小的顺序排列的序号, $1 \leq s \leq DN$ , $\text{SqCoef}_s$ 为在所述系数序列中排序为第 $s$ 位的权重系数;

从所述系数序列中选取排列在前G位的权重系数,且G的取值满足下列条件:

$$\begin{cases} \sum_{s=1}^G SqCoef_s > Scale \\ \sum_{s=1}^{G-1} SqCoef_s \leq Scale \end{cases}$$

其中,Scale为预设的比例阈值,且 $0 < Scale < 1$ 。

4. 根据权利要求2所述的用戶数据分析方法,其特征在于,所述使用所述样本集合对所述数据分析模型进行训练,求解出所述第一模型系数向量包括:

根据下式计算对所述数据分析模型进行第t次训练的误差向量:

$$Error_t = \begin{bmatrix} TgVal_1 \\ TgVal_2 \\ \dots \\ TgVal_u \\ \dots \\ TgVal_{UN} \end{bmatrix} - Sigmoid \left( \begin{bmatrix} UserVec_1 \\ UserVec_2 \\ \dots \\ UserVec_u \\ \dots \\ UserVec_{UN} \end{bmatrix} \times CoefVec_t^T \right)$$

其中,CoefVec<sub>t</sub>为第t次训练的第一模型系数向量,且CoefVec<sub>t</sub> = (Coef<sub>1,t</sub>, Coef<sub>2,t</sub>, ..., Coef<sub>d,t</sub>, ..., Coef<sub>DN,t</sub>), Coef<sub>d,t</sub>为第d个分析维度上的权重系数在第t次训练中的取值, TgVal<sub>u</sub>为所述样本集合中第u个历史用户的目标结果, Error<sub>t</sub>为对所述数据分析模型进行第t次训练的误差向量,且Error<sub>t</sub> = (ErrEm<sub>t,1</sub>, ErrEm<sub>t,2</sub>, ..., ErrEm<sub>t,u</sub>, ..., ErrEm<sub>t,UN</sub>)<sup>T</sup>, ErrEm<sub>t,un</sub>为所述样本集合中第u个历史用户在第t次训练中的误差值;

根据第t次训练的第一模型系数向量以及第t次训练的误差向量计算第t+1次训练的第一模型系数向量;

根据下式计算第t次训练的模型系数向量距离:

$$Dis_t = |CoefVec_{t+1} - CoefVec_t| = \sqrt{\sum_{d=1}^{DN} (Coef_{d,t+1} - Coef_{d,t})^2}$$

其中,Dis<sub>t</sub>为第t次训练的模型系数向量距离;

若第t次训练的模型系数向量距离大于预设的距离阈值,则对所述数据分析模型进行第t+1次训练;

若第t次训练的模型系数向量距离小于或等于所述距离阈值,则结束对所述数据分析模型的训练。

5. 根据权利要求4所述的用戶数据分析方法,其特征在于,所述根据第t次训练的第一模型系数向量以及第t次训练的误差向量计算第t+1次训练的第一模型系数向量包括:

根据下式计算第t+1次训练的第一模型系数向量:

$$CoefVec_{t+1} = CoefVec_t - (\partial \times \begin{bmatrix} UserVec_1 \\ UserVec_2 \\ \dots \\ UserVec_u \\ \dots \\ UserVec_{UN} \end{bmatrix} \times Error_t)^T$$

其中,  $\partial$  为预设的迭代步长。

6. 一种用户数据分析装置, 其特征在于, 包括:

分析请求接收模块, 用于接收终端设备发送的用户数据分析请求, 并从所述用户数据分析请求中提取待分析用户的身份标识;

用户数据查询模块, 用于根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据;

用户特征向量构造模块, 用于根据各个分析维度上的用户数据构造所述待分析用户的用户特征向量;

用户数据分析模块, 用于使用预设的数据分析模型对所述待分析用户的用户特征向量进行计算, 得到所述待分析用户的分析结果, 所述数据分析模型为经过预设的样本集合训练的机器学习模型;

交互操作模块, 用于根据所述分析结果执行与所述待分析用户的交互操作。

7. 根据权利要求6所述的装置, 其特征在于, 还包括:

数据分析模型构建模块, 用于根据下式构建所述数据分析模型:

$$CalRes_u = \text{Sigmoid}(CoefVec \times UserVec_u^T)$$

其中, Sigmoid 为 S 型生长曲线函数,  $u$  为所述样本集合中的历史用户的序号,  $1 \leq u \leq UN$ ,  $UserVec_u$  为所述样本集合中第  $u$  个历史用户的用户特征向量, 且  $UserVec_u = (UsData_{u,1}, UsData_{u,2}, \dots, UsData_{u,d}, \dots, UsData_{u,DN})$ ,  $UsData_{u,d}$  为所述样本集合中第  $u$  个历史用户在第  $d$  个分析维度上的用户数据,  $CoefVec$  为预设的第一模型系数向量, 且  $CoefVec = (Coef_1, Coef_2, \dots, Coef_d, \dots, Coef_{DN})$ ,  $Coef_d$  为第  $d$  个分析维度上的权重系数,  $T$  为转置符号,  $CalRes_u$  为所述样本集合中第  $u$  个历史用户的模型输出值;

第一模型训练模块, 用于使用所述样本集合对所述数据分析模型进行训练, 求解出所述第一模型系数向量;

权重系数选取模块, 用于从所述第一模型系数向量中选取  $G$  个权重系数, 并构建如下所示的第二模型系数向量:

$$NwCoefVec = (SqCoef_1, SqCoef_2, \dots, SqCoef_{nd}, \dots, SqCoef_G)$$

其中,  $NwCoefVec$  为所述第二模型系数向量,  $nd$  为所述第二模型系数向量中的各个权重系数的序号,  $1 \leq nd \leq G$ ,  $G$  为正整数,  $SqCoef_{nd}$  为所述第二模型系数向量中的第  $nd$  个权重系数;

数据分析模型更新模块, 用于根据下式对所述数据分析模型进行更新:

$$CalRes_u = \text{Sigmoid}(NwCoefVec \times NwUserVec_u^T)$$

其中,  $NwUserVec_u = (NwUsData_{u,1}, NwUsData_{u,2}, \dots, NwUsData_{u,nd}, \dots, NwUsData_{u,G})$ ,  $NwUsData_{u,s}$  为所述样本集合中第  $u$  个历史用户与所述第二模型系数向量中的第  $nd$  个权重系数对应的分析维度上的用户数据;

第二模型训练模块, 用于使用所述样本集合对更新后的所述数据分析模型进行训练, 求解出所述第二模型系数向量。

8. 根据权利要求7所述的用于用户数据分析方法, 其特征在于, 所述权重系数选取模块包括:

系数序列排列单元, 用于将所述第一模型系数向量中的各个权重系数按照从大到小的顺序排列为如下所示的系数序列:

$$SqCoef_1, SqCoef_2, \dots, SqCoef_s, \dots, SqCoef_{DN}$$

其中,  $s$  为所述第一模型系数向量中的各个权重系数按照从大到小的顺序排列的序号,  $1 \leq s \leq DN$ ,  $SqCoef_s$  为在所述系数序列中排序为第  $s$  位的权重系数;

权重系数选取单元, 用于从所述系数序列中选取排列在前  $G$  位的权重系数, 且  $G$  的取值满足下列条件:

$$\begin{cases} \sum_{s=1}^G SqCoef_s > Scale \\ \sum_{s=1}^{G-1} SqCoef_s \leq Scale \end{cases}$$

其中,  $Scale$  为预设的比例阈值, 且  $0 < Scale < 1$ 。

9. 一种计算机可读存储介质, 所述计算机可读存储介质存储有计算机可读指令, 其特征在于, 所述计算机可读指令被处理器执行时实现如权利要求1至5中任一项所述的用户数据分析方法的步骤。

10. 一种终端设备, 包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令, 其特征在于, 所述处理器执行所述计算机可读指令时实现如权利要求1至5中任一项所述的用户数据分析方法的步骤。

## 一种用户数据分析方法、装置、可读存储介质及终端设备

### 技术领域

[0001] 本发明属于计算机技术领域,尤其涉及一种用户数据分析方法、装置、计算机可读存储介质及终端设备。

### 背景技术

[0002] 随着移动互联网技术的不断发展,银行、信贷公司等金融机构往往可以获取海量的用户数据,这些数据与用户对金融产品或者服务的偏好往往具有较强的内在关联,而现有技术中并未对这些用户数据进行充分的挖掘分析,无法及时根据用户的需求与用户进行对应的交互,导致用户体验较差。

### 发明内容

[0003] 有鉴于此,本发明实施例提供了一种用户数据分析方法、装置、计算机可读存储介质及终端设备,以解决现有技术中并未对这些用户数据进行充分的挖掘分析,无法及时根据用户的需求与用户进行对应的交互,导致用户体验较差的问题。

[0004] 本发明实施例的第一方面提供了一种用户数据分析方法,可以包括:

[0005] 接收终端设备发送的用户数据分析请求,并从所述用户数据分析请求中提取待分析用户的身份标识;

[0006] 根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据;

[0007] 根据各个分析维度上的用户数据构造所述待分析用户的用户特征向量;

[0008] 使用预设的数据分析模型对所述待分析用户的用户特征向量进行计算,得到所述待分析用户的分析结果,所述数据分析模型为经过预设的样本集合训练的机器学习模型;

[0009] 根据所述分析结果执行与所述待分析用户的交互操作。

[0010] 本发明实施例的第二方面提供了一种用户数据分析装置,可以包括:

[0011] 分析请求接收模块,用于接收终端设备发送的用户数据分析请求,并从所述用户数据分析请求中提取待分析用户的身份标识;

[0012] 用户数据查询模块,用于根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据;

[0013] 用户特征向量构造模块,用于根据各个分析维度上的用户数据构造所述待分析用户的用户特征向量;

[0014] 用户数据分析模块,用于使用预设的数据分析模型对所述待分析用户的用户特征向量进行计算,得到所述待分析用户的分析结果,所述数据分析模型为经过预设的样本集合训练的机器学习模型;

[0015] 交互操作模块,用于根据所述分析结果执行与所述待分析用户的交互操作。

[0016] 本发明实施例的第三方面提供了一种计算机可读存储介质,所述计算机可读存储介质存储有计算机可读指令,所述计算机可读指令被处理器执行时实现如下步骤:

[0017] 接收终端设备发送的用户数据分析请求,并从所述用户数据分析请求中提取待分析用户的身份标识;

[0018] 根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据;

[0019] 根据各个分析维度上的用户数据构造所述待分析用户的用户特征向量;

[0020] 使用预设的数据分析模型对所述待分析用户的用户特征向量进行计算,得到所述待分析用户的分析结果,所述数据分析模型为经过预设的样本集合训练的机器学习模型;

[0021] 根据所述分析结果执行与所述待分析用户的交互操作。

[0022] 本发明实施例的第四方面提供了一种终端设备,包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令,所述处理器执行所述计算机可读指令时实现如下步骤:

[0023] 接收终端设备发送的用户数据分析请求,并从所述用户数据分析请求中提取待分析用户的身份标识;

[0024] 根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据;

[0025] 根据各个分析维度上的用户数据构造所述待分析用户的用户特征向量;

[0026] 使用预设的数据分析模型对所述待分析用户的用户特征向量进行计算,得到所述待分析用户的分析结果,所述数据分析模型为经过预设的样本集合训练的机器学习模型;

[0027] 根据所述分析结果执行与所述待分析用户的交互操作。

[0028] 本发明实施例与现有技术相比存在的有益效果是:本发明实施例预先构建起对用户数据进行分析的机器学习模型,并通过预设的样本集合完成对其的训练过程。在接收到终端设备发送的用户数据分析请求时,首先从所述用户数据分析请求中提取出待分析用户的身份标识,然后根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据,据此构造出所述待分析用户的用户特征向量,并使用训练好的数据分析模型对所述待分析用户的用户特征向量进行计算,得到所述待分析用户的分析结果,最后根据所述分析结果执行与所述待分析用户的交互操作。即通过使用机器学习的方法对用户数据进行充分的挖掘分析,及时了解用户的需求,并根据用户的需求与用户进行对应的交互,以尽量满足用户的需求,从而极大提升了用户的体验。

## 附图说明

[0029] 为了更清楚地说明本发明实施例中的技术方案,下面将对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发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其它的附图。

[0030] 图1为本发明实施例中一种用户数据分析方法的一个实施例流程图;

[0031] 图2为数据分析模型的构建过程的示意图;

[0032] 图3为使用样本集合对数据分析模型进行训练的示意图;

[0033] 图4为本发明实施例中一种用户数据分析装置的一个实施例结构图;

[0034] 图5为本发明实施例中一种终端设备的示意框图。

## 具体实施方式

[0035] 为使得本发明的发明目的、特征、优点能够更加的明显和易懂,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,下面所描述的实施例仅仅是本发明一部分实施例,而非全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例,都属于本发明保护的范围。

[0036] 请参阅图1,本发明实施例中一种用户数据分析方法的一个实施例可以包括:

[0037] 步骤S101、接收终端设备发送的用户数据分析请求,并从所述用户数据分析请求中提取待分析用户的身份标识。

[0038] 所述身份标识包括但不限于待分析用户的身份证号码、手机号码、社保号码、驾照号码、公积金号码或者其它可以唯一表征用户身份的标识。

[0039] 步骤S102、根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据。

[0040] 在所述数据库中包括了各个用户的用户数据,每个用户的用户数据均以该用户的身份标识作为索引,这些用户数据包括但不限于以下分析维度:

[0041] 用户最近一次购买金融产品的时间距离目前的天数:缺失赋值382,变量值范围0~6049,最近一次购买过平安产品的时间距今越近,越倾向于成为贷款用户。

[0042] 用户最近一年的信用卡月均账单金额:缺失赋值0,变量值范围:0~43987,信用卡账单月均余额越高,则说明用户消费能力越强,越倾向于成为贷款用户。

[0043] 用户持有金融产品的合同数:缺失赋值0,变量值范围0~17,用户持有的金融产品数越多,越倾向于成为贷款用户。

[0044] 用户价值等级:缺失赋值0,变量值范围0~4,即共有5个价值等级,变量取值越大,则用户价值等级也越高,用户价值越高的用户越倾向于成为贷款用户。

[0045] 用户房产均价:缺失赋值9444,变量值范围:0~67315,房产均价越高的用户越倾向于成为贷款用户。

[0046] 用户职业等级:缺失赋值0,变量值范围0~4,即共有5个职业等级,变量取值越大,则用户职业等级也越高,用户职业等级越高的用户越倾向于成为贷款用户。

[0047] 是否有车:缺失赋值0,变量值范围0~1,其中,0代表无车,1代表有车,有车的用户比较倾向于成为贷款用户。

[0048] 用户在汽车购买、保养和维修以及在家电、装修等商户的消费金额:缺失赋值0,变量值范围0~1000000,消费集中在汽车领域和家电装修的用户倾向于成为贷款用户。

[0049] 用户最近一年交易活跃次数:缺失赋值0,变量值范围0~356,交易活跃次数越多的用户比较倾向于成为贷款用户。

[0050] 用户最近一年合计有效及失效险种实缴保费:缺失赋值0,变量值范围0~84615,最近一年保费越多的用户越倾向于成为贷款用户。

[0051] 用户是否持有寿险意外医疗保险产品:缺失赋值0,变量值范围0~1,其中,0代表未持有寿险意外医疗保险产品,1代表持有寿险意外医疗保险产品,持有寿险意外医疗保险产品的用户倾向于成为贷款用户。

[0052] 在提取出所述待分析用户的身份标识之后,即可以所述待分析用户的身份标识作

为索引从所述数据库中查询出所述待分析用户在上述各个分析维度上的用户数据。

[0053] 步骤S103、根据各个分析维度上的用户数据构造所述待分析用户的用户特征向量。

[0054] 具体地,可以构造如下式所示的用户特征向量:

[0055]  $TgVec = (TgData_1, TgData_2, \dots, TgData_d, \dots, TgData_{DN})$

[0056] 其中, $d$ 为各个分析维度的序号, $1 \leq d \leq DN$ , $DN$ 为分析维度的总数, $TgData_d$ 为所述待分析用户在第 $d$ 个分析维度上的用户数据, $TgVec$ 为所述待分析用户的用户特征向量。

[0057] 步骤S104、使用预设的数据分析模型对所述待分析用户的用户特征向量进行计算,得到所述待分析用户的分析结果。

[0058] 所述数据分析模型为经过预设的样本集合训练的机器学习模型。

[0059] 为了构造出一个准确的数据分析模型,需要通过大量历史用户的用户数据来对模型进行训练,每个历史用户的用户数据均可构造成一个对应的用户特征向量,这些用户特征向量构成了所述样本集合。本实施例中按照历史用户最终是否申请贷款将样本集合中的历史用户划分为正向用户和负向用户两类,其中,正向用户为目标结果为预设的第一结果(取值为1,即已申请贷款)的历史用户,负向用户为目标结果为预设的第二结果(取值为0,即未申请贷款)的历史用户。

[0060] 为了尽量保证数据分析模型具有广泛的适用性,在选取样本时,需要尽量保证正向用户与负向用户之间的均衡性,以及各个分析维度数据分布的均衡性,在从所述数据库中选取历史用户的用户特征向量组成所述样本集合时,选取的历史用户需要满足以下条件:

[0061] 选取满足以下条件的 $UN$ 个历史用户的用户特征向量组成所述样本集合:

$$[0062] \begin{cases} \frac{|PUN - NUN|}{UN} < DiffThresh1 \\ \frac{\sum_{v=1}^{VN_d} (CountNum_{d,v} - \frac{UN}{VN_d})^2}{VN_d} < DiffThresh2 \end{cases}$$

[0063] 其中, $UN$ 为从所述数据库中选取的历史用户的数目,且 $UN$ 为正整数, $PUN$ 为所述样本集合中的正向用户的数目, $NUN$ 为所述样本集合中的负向用户的数目,且 $UN = PUN + NUN$ , $v$ 为各个分析维度上的取值的序号, $1 \leq v \leq VN_d$ , $VN_d$ 为在第 $d$ 个分析维度上的取值的总数,例如,对于用户持有金融产品的合同数这一分析维度而言,共有0、1、2、3、...、17这18个取值,则在这一维度上,有 $VN_d = 18$ , $CountNum_{d,v}$ 为在第 $d$ 个分析维度上的取值为第 $v$ 个取值的历史用户的数目, $DiffThresh1$ 为预设的第一阈值,其具体取值可以根据实际情况进行设置,例如,可以将其设置为0.1、0.01、0.001或者其它取值, $DiffThresh2$ 为预设的第二阈值,其具体取值可以根据实际情况进行设置,例如,可以将其设置为20、50、100或者其它取值。

[0064] 在本实施例中,可以将所述样本集合中的各个历史用户的用户特征向量记为:

[0065]  $UserVec_u = (UsData_{u,1}, UsData_{u,2}, \dots, UsData_{u,d}, \dots, UsData_{u, DN})$

[0066] 其中, $u$ 为所述样本集合中的历史用户的序号, $1 \leq u \leq UN$ , $UserVec_u$ 为所述样本集合中第 $u$ 个历史用户的用户特征向量, $UsData_{u,d}$ 为所述样本集合中第 $u$ 个历史用户在第 $d$ 个

分析维度上的用户数据。

[0067] 在完成了以上的对所述样本集合的构建过程之后,可以通过如图2所示的过程构建所述数据分析模型:

[0068] 步骤S201、根据下式构建所述数据分析模型:

$$[0069] \quad \text{CalRes}_u = \text{Sigmoid}(\text{CoefVec} \times \text{UserVec}_u^T)$$

[0070] 其中,Sigmoid为S型生长曲线函数,CoefVec为预设的第一模型系数向量,且 $\text{CoefVec} = (\text{Coef}_1, \text{Coef}_2, \dots, \text{Coef}_d, \dots, \text{Coef}_{DN})$ , $\text{Coef}_d$ 为第d个分析维度上的权重系数,T为转置符号,CalRes<sub>u</sub>为所述样本集合中第u个历史用户的模型输出值。

[0071] 步骤S202、使用所述样本集合对所述数据分析模型进行训练,求解出所述第一模型系数向量。

[0072] 对所述数据分析模型进行训练的过程是一个循环迭代的过程,以下以图3所示的第t次训练过程为例对其进行详细说明:

[0073] 步骤S2021、计算对所述数据分析模型进行第t次训练的误差向量。

[0074] 具体地,可以根据下式计算对所述数据分析模型进行第t次训练的误差向量:

$$[0075] \quad \text{Error}_t = \begin{bmatrix} \text{TgVal}_1 \\ \text{TgVal}_2 \\ \dots \\ \text{TgVal}_u \\ \dots \\ \text{TgVal}_{UN} \end{bmatrix} - \text{Sigmoid} \left( \begin{bmatrix} \text{UserVec}_1 \\ \text{UserVec}_2 \\ \dots \\ \text{UserVec}_u \\ \dots \\ \text{UserVec}_{UN} \end{bmatrix} \times \text{CoefVec}_t^T \right)$$

[0076] 其中,CoefVec<sub>t</sub>为第t次训练的第一模型系数向量,且 $\text{CoefVec}_t = (\text{Coef}_{1,t}, \text{Coef}_{2,t}, \dots, \text{Coef}_{d,t}, \dots, \text{Coef}_{DN,t})$ , $\text{Coef}_{d,t}$ 为第d个分析维度上的权重系数在第t次训练中的取值,,特殊地,在初始状态下(也即在第1次训练时),可以将所述第一模型系数向量设置为取值为全1的向量,即: $\text{CoefVec}_1 = (1, 1, \dots, 1, \dots, 1)$ ,TgVal<sub>u</sub>为所述样本集合中第u个历史用户的目标结果,Error<sub>t</sub>为对所述数据分析模型进行第t次训练的误差向量,且 $\text{Error}_t = (\text{ErrEm}_{t,1}, \text{ErrEm}_{t,2}, \dots, \text{ErrEm}_{t,u}, \dots, \text{ErrEm}_{t,UN})^T$ ,ErrEm<sub>t,un</sub>为所述样本集合中第u个历史用户在第t次训练中的误差值。

[0077] 步骤S2022、根据第t次训练的第一模型系数向量以及第t次训练的误差向量计算第t+1次训练的第一模型系数向量。

[0078] 具体地,可以根据下式计算第t+1次训练的第一模型系数向量:

$$[0079] \quad \text{CoefVec}_{t+1} = \text{CoefVec}_t - (\partial \times \begin{bmatrix} \text{UserVec}_1 \\ \text{UserVec}_2 \\ \dots \\ \text{UserVec}_u \\ \dots \\ \text{UserVec}_{UN} \end{bmatrix} \times \text{Error}_t)^T$$

[0080] 其中,  $\partial$  为预设的迭代步长, 也即迭代计算过程中每次迭代的速率, 就相当于爬山中每步的跨度, 如果步长太长, 在训练过程中可能一下子就迈过了最合适的点, 失去了最优解, 如果步长太短, 则会发现达到终点的速率会非常慢, 容易造成过拟合。在本实施例中, 可以根据实际情况对迭代步长进行设置, 例如, 可以将其设置为0.1、0.01、0.001或者其它取值。

[0081] 步骤S2023、计算第t次训练的模型系数向量距离。

[0082] 具体地, 可以根据下式计算第t次训练的模型系数向量距离:

$$[0083] \quad Dis_t = |CoefVec_{t+1} - CoefVec_t| = \sqrt{\sum_{d=1}^{DN} (Coef_{d,t+1} - Coef_{d,t})^2}$$

[0084] 其中,  $Dis_t$  为第t次训练的模型系数向量距离。

[0085] 步骤S2024、判断第t次训练的模型系数向量距离是否大于预设的距离阈值。

[0086] 所述距离阈值可以根据实际情况进行设置, 例如, 可以将其设置为5、10、20或者其它取值。

[0087] 若第t次训练的模型系数向量距离大于所述距离阈值, 则执行步骤S2025, 若第t次训练的模型系数向量距离小于或等于所述距离阈值, 则执行步骤S2026。

[0088] 步骤S2025、对所述数据分析模型进行第t+1次训练。

[0089] 第t+1次训练的具体过程与图2所示的第t次训练过程类似, 此处不再赘述。

[0090] 步骤S2026、结束对所述数据分析模型的训练。

[0091] 经过上述过程, 可得到最终的第一模型系数向量, 也即训练结束时的第一模型系数向量。但是, 在前期进行用户数据采集时, 为了保证数据的全面性, 往往会采集众多分析维度的用户数据, 也即DN的值会极大, 各个分析维度的用户数据对最终结果的影响各不相同, 有的较大, 有的较小, 因此, 可以根据上一步的计算结果从中剔除掉对结果影响不大的分析维度上的数据, 从而减少后续对用户数据进行分析的计算量。

[0092] 步骤S203、从所述第一模型系数向量中选取G个权重系数。

[0093] 具体地, 首先将所述第一模型系数向量中的各个权重系数按照从大到小的顺序排列为如下所示的系数序列:

[0094]  $SqCoef_1, SqCoef_2, \dots, SqCoef_s, \dots, SqCoef_{DN}$

[0095] 其中, s为所述第一模型系数向量中的各个权重系数按照从大到小的顺序排列的序号,  $1 \leq s \leq DN$ ,  $SqCoef_s$  为在所述系数序列中排序为第s位的权重系数;

[0096] 然后, 从所述系数序列中选取排列在前G位的权重系数, G为正整数, 且G的取值满足下列条件:

$$[0097] \quad \begin{cases} \sum_{s=1}^G SqCoef_s > Scale \\ \sum_{s=1}^{G-1} SqCoef_s \leq Scale \end{cases}$$

[0098] 其中, Scale为预设的比例阈值, 且  $0 < Scale < 1$ , 其具体取值可以根据实际情况进行设置, 例如, 可以将其设置为98%、95%、90%或者其它取值。

[0099] 则在后续的计算中,只选取与这G个优选系数对应的分析维度上的用户数据,而剔除掉其它分析维度上的用户数据,从而达到简化运算的目的。

[0100] 优选地,由于剔除掉了很多对结果影响较小的权重系数,剩余的优选系数的具体取值也会随之发生变化,因此可以重新对选取出的优选系数再次进行模型训练,对其取值进行调整。

[0101] 步骤S204、构建如下所示的第二模型系数向量:

[0102]  $NwCoefVec = (SqCoef_1, SqCoef_2, \dots, SqCoef_{nd}, \dots, SqCoef_G)$

[0103] 其中, $NwCoefVec$ 为所述第二模型系数向量, $nd$ 为所述第二模型系数向量中的各个权重系数的序号, $1 \leq nd \leq G$ , $SqCoef_{nd}$ 为所述第二模型系数向量中的第 $nd$ 个权重系数。

[0104] 步骤S205、根据下式对所述数据分析模型进行更新:

[0105]  $CalRes_u = Sigmoid(NwCoefVec \times NwUserVec_u^T)$

[0106] 其中, $NwUserVec_u = (NwUsData_{u,1}, NwUsData_{u,2}, \dots, NwUsData_{u,nd}, \dots, NwUsData_{u,G})$ , $NwUsData_{u,s}$ 为所述样本集合中第 $u$ 个历史用户与所述第二模型系数向量中的第 $nd$ 个权重系数对应的分析维度上的用户数据。

[0107] 步骤S206、使用所述样本集合对更新后的所述数据分析模型进行训练,求解出所述第二模型系数向量。

[0108] 在完成了数据分析模型的更新之后,再次使用所述样本集合对其进行训练,具体的训练过程与前述的训练过程类似,具体可参照前述详细描述,此处不再赘述。

[0109] 在数据分析模型训练完成之后,即可使用该数据分析模型对所述待分析用户的用户特征向量进行计算,得到一个输出值,也即所述待分析用户的分析结果。

[0110] 步骤S105、根据所述分析结果执行与所述待分析用户的交互操作。

[0111] 若所述待分析用户的分析结果小于预设的分析阈值,则可认为其申请贷款的几率较小,不再执行与其的任何交互操作,若所述待分析用户的分析结果大于所述分析阈值,则可认为其申请贷款的几率较大,此时则可以通过发送营销短信等手段与其进行交互操作。所述分析阈值可以根据实际情况进行设置,例如,可以将其设置为0.4、0.5、0.6或者其它取值,在本实施例中,优选将其设置为0.5。

[0112] 综上所述,本发明实施例预先构建起对用户数据进行分析的机器学习模型,并通过预设的样本集合完成对其的训练过程。在接收到终端设备发送的用户数据分析请求时,首先从所述用户数据分析请求中提取出待分析用户的身份标识,然后根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据,据此构造出所述待分析用户的用户特征向量,并使用训练好的数据分析模型对所述待分析用户的用户特征向量进行计算,得到所述待分析用户的分析结果,最后根据所述分析结果执行与所述待分析用户的交互操作。即通过使用机器学习的方法对用户数据进行充分的挖掘分析,及时了解用户的需求,并根据用户的需求与用户进行对应的交互,以尽量满足用户的需求,从而极大提升了用户的体验。

[0113] 应理解,上述实施例中各步骤的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定,而不应对本发明实施例的实施过程构成任何限定。

[0114] 对应于上文实施例所述的一种用户数据分析方法,图4示出了本发明实施例提供

的一种用户数据分析装置的一个实施例结构图。

[0115] 本实施例中,一种用户数据分析装置可以包括:

[0116] 分析请求接收模块401,用于接收终端设备发送的用户数据分析请求,并从所述用户数据分析请求中提取待分析用户的身份标识;

[0117] 用户数据查询模块402,用于根据所述待分析用户的身份标识从预设的数据库中查询所述待分析用户在预设的各个分析维度上的用户数据;

[0118] 用户特征向量构造模块403,用于根据各个分析维度上的用户数据构造所述待分析用户的用户特征向量;

[0119] 用户数据分析模块404,用于使用预设的数据分析模型对所述待分析用户的用户特征向量进行计算,得到所述待分析用户的分析结果,所述数据分析模型为经过预设的样本集合训练的机器学习模型;

[0120] 交互操作模块405,用于根据所述分析结果执行与所述待分析用户的交互操作。

[0121] 进一步地,所述用户数据分析装置还可以包括:

[0122] 数据分析模型构建模块,用于根据下式构建所述数据分析模型:

$$[0123] \quad \text{CalRes}_u = \text{Sigmoid}(\text{CoefVec} \times \text{UserVec}_u^T)$$

[0124] 其中,Sigmoid为S型生长曲线函数,u为所述样本集合中的历史用户的序号, $1 \leq u \leq UN$ , $\text{UserVec}_u$ 为所述样本集合中第u个历史用户的用户特征向量,且 $\text{UserVec}_u = (\text{UsData}_{u,1}, \text{UsData}_{u,2}, \dots, \text{UsData}_{u,d}, \dots, \text{UsData}_{u,DN})$ , $\text{UsData}_{u,d}$ 为所述样本集合中第u个历史用户在第d个分析维度上的用户数据,CoefVec为预设的第一模型系数向量,且 $\text{CoefVec} = (\text{Coef}_1, \text{Coef}_2, \dots, \text{Coef}_d, \dots, \text{Coef}_{DN})$ , $\text{Coef}_d$ 为第d个分析维度上的权重系数,T为转置符号,CalRes<sub>u</sub>为所述样本集合中第u个历史用户的模型输出值;

[0125] 第一模型训练模块,用于使用所述样本集合对所述数据分析模型进行训练,求解出所述第一模型系数向量;

[0126] 权重系数选取模块,用于从所述第一模型系数向量中选取G个权重系数,并构建如下所示的第二模型系数向量:

$$[0127] \quad \text{NwCoefVec} = (\text{SqCoef}_1, \text{SqCoef}_2, \dots, \text{SqCoef}_{nd}, \dots, \text{SqCoef}_G)$$

[0128] 其中,NwCoefVec为所述第二模型系数向量,nd为所述第二模型系数向量中的各个权重系数的序号, $1 \leq nd \leq G$ ,G为正整数,SqCoef<sub>nd</sub>为所述第二模型系数向量中的第nd个权重系数;

[0129] 数据分析模型更新模块,用于根据下式对所述数据分析模型进行更新:

$$[0130] \quad \text{CalRes}_u = \text{Sigmoid}(\text{NwCoefVec} \times \text{NwUserVec}_u^T)$$

[0131] 其中,NwUserVec<sub>u</sub> = (NwUsData<sub>u,1</sub>, NwUsData<sub>u,2</sub>, ..., NwUsData<sub>u,nd</sub>, ..., NwUsData<sub>u,G</sub>),NwUsData<sub>u,s</sub>为所述样本集合中第u个历史用户与所述第二模型系数向量中的第nd个权重系数对应的分析维度上的用户数据;

[0132] 第二模型训练模块,用于使用所述样本集合对更新后的所述数据分析模型进行训练,求解出所述第二模型系数向量。

[0133] 进一步地,所述权重系数选取模块可以包括:

[0134] 系数序列排列单元,用于将所述第一模型系数向量中的各个权重系数按照从大到小的顺序排列为如下所示的系数序列:

[0135]  $SqCoef_1, SqCoef_2, \dots, SqCoef_s, \dots, SqCoef_{DN}$

[0136] 其中,  $s$ 为所述第一模型系数向量中的各个权重系数按照从大到小的顺序排列的序号,  $1 \leq s \leq DN$ ,  $SqCoef_s$ 为在所述系数序列中排序为第 $s$ 位的权重系数;

[0137] 权重系数选取单元,用于从所述系数序列中选取排列在前 $G$ 位的权重系数,且 $G$ 的取值满足下列条件:

$$[0138] \quad \begin{cases} \sum_{s=1}^G SqCoef_s > Scale \\ \sum_{s=1}^{G-1} SqCoef_s \leq Scale \end{cases}$$

[0139] 其中,  $Scale$ 为预设的比例阈值,且 $0 < Scale < 1$ 。

[0140] 进一步地,所述第一模型训练模块可以包括:

[0141] 误差向量计算单元,用于根据下式计算对所述数据分析模型进行第 $t$ 次训练的误差向量:

$$[0142] \quad Error_t = \begin{bmatrix} TgVal_1 \\ TgVal_2 \\ \dots \\ TgVal_u \\ \dots \\ TgVal_{UN} \end{bmatrix} - Sigmoid \left( \begin{bmatrix} UserVec_1 \\ UserVec_2 \\ \dots \\ UserVec_u \\ \dots \\ UserVec_{UN} \end{bmatrix} \times CoefVec_t^T \right)$$

[0143] 其中,  $CoefVec_t$ 为第 $t$ 次训练的第一模型系数向量,且 $CoefVec_t = (Coef_{1,t}, Coef_{2,t}, \dots, Coef_{d,t}, \dots, Coef_{DN,t})$ ,  $Coef_{d,t}$ 为第 $d$ 个分析维度上的权重系数在第 $t$ 次训练中的取值,  $TgVal_u$ 为所述样本集合中第 $u$ 个历史用户的目标结果,  $Error_t$ 为对所述数据分析模型进行第 $t$ 次训练的误差向量,且 $Error_t = (ErrEm_{t,1}, ErrEm_{t,2}, \dots, ErrEm_{t,u}, \dots, ErrEm_{t,UN})^T$ ,  $ErrEm_{t,un}$ 为所述样本集合中第 $u$ 个历史用户在第 $t$ 次训练中的误差值;

[0144] 第一模型系数向量计算单元,用于根据第 $t$ 次训练的第一模型系数向量以及第 $t$ 次训练的误差向量计算第 $t+1$ 次训练的第一模型系数向量;

[0145] 向量距离计算单元,用于根据下式计算第 $t$ 次训练的模型系数向量距离:

$$[0146] \quad Dis_t = |CoefVec_{t+1} - CoefVec_t| = \sqrt{\sum_{d=1}^{DN} (Coef_{d,t+1} - Coef_{d,t})^2}$$

[0147] 其中,  $Dis_t$ 为第 $t$ 次训练的模型系数向量距离;

[0148] 第一处理模块,用于若第 $t$ 次训练的模型系数向量距离大于预设的距离阈值,则对所述数据分析模型进行第 $t+1$ 次训练;

[0149] 第二处理模块,用于若第 $t$ 次训练的模型系数向量距离小于或等于所述距离阈值,则结束对所述数据分析模型的训练。

[0150] 进一步地,所述第一模型系数向量计算模块具体用于根据下式计算第 $t+1$ 次训练的第一模型系数向量:

$$[0151] \quad \text{CoefVec}_{t+1} = \text{CoefVec}_t - (\partial \times \begin{bmatrix} \text{UserVec}_1 \\ \text{UserVec}_2 \\ \dots \\ \text{UserVec}_u \\ \dots \\ \text{UserVec}_{UN} \end{bmatrix} \times \text{Error}_t)^T$$

[0152] 其中， $\partial$  为预设的迭代步长。

[0153] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的装置，模块和单元的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0154] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中未详述或记载的部分，可以参见其它实施例的相关描述。

[0155] 图5示出了本发明实施例提供的一种终端设备的示意框图，为了便于说明，仅示出了与本发明实施例相关的部分。

[0156] 在本实施例中，所述终端设备5可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。该终端设备5可包括：处理器50、存储器51以及存储在所述存储器51中并可在所述处理器50上运行的计算机可读指令52，例如执行上述的用户数据分析方法的计算机可读指令。所述处理器50执行所述计算机可读指令52时实现上述各个用户数据分析方法实施例中的步骤，例如图1所示的步骤S101至S105。或者，所述处理器50执行所述计算机可读指令52时实现上述各装置实施例中各模块/单元的功能，例如图4所示模块401至405的功能。

[0157] 示例性的，所述计算机可读指令52可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器51中，并由所述处理器50执行，以完成本发明。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机可读指令段，该指令段用于描述所述计算机可读指令52在所述终端设备5中的执行过程。

[0158] 所述处理器50可以是中央处理单元 (Central Processing Unit, CPU)，还可以是其它通用处理器、数字信号处理器 (Digital Signal Processor, DSP)、专用集成电路 (Application Specific Integrated Circuit, ASIC)、现场可编程门阵列 (Field-Programmable Gate Array, FPGA) 或者其它可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0159] 所述存储器51可以是所述终端设备5的内部存储单元，例如终端设备5的硬盘或内存。所述存储器51也可以是所述终端设备5的外部存储设备，例如所述终端设备5上配备的插接式硬盘，智能存储卡 (Smart Media Card, SMC)，安全数字 (Secure Digital, SD) 卡，闪存卡 (Flash Card) 等。进一步地，所述存储器51还可以既包括所述终端设备5的内部存储单元也包括外部存储设备。所述存储器51用于存储所述计算机可读指令以及所述终端设备5所需的其它指令和数据。所述存储器51还可以用于暂时地存储已经输出或者将要输出的数据。

[0160] 在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。

[0161] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干计算机可读指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储计算机可读指令的介质。

[0162] 以上所述实施例仅用以说明本发明的技术方案,而非对其限制;尽管参照前述实施例对本发明进行了详细的说明,本领域的普通技术人员应当理解:其依然可以对前述各实施例所记载的技术方案进行修改,或者对其中部分技术特征进行等同替换;而这些修改或者替换,并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

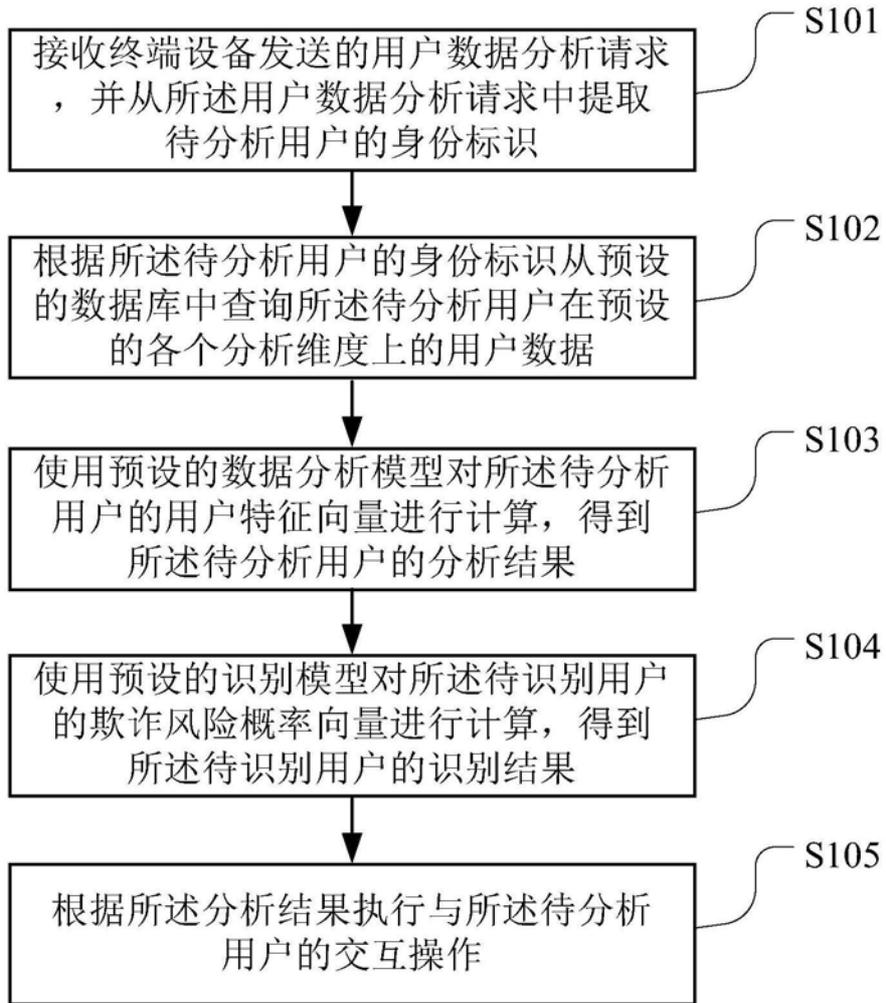


图1

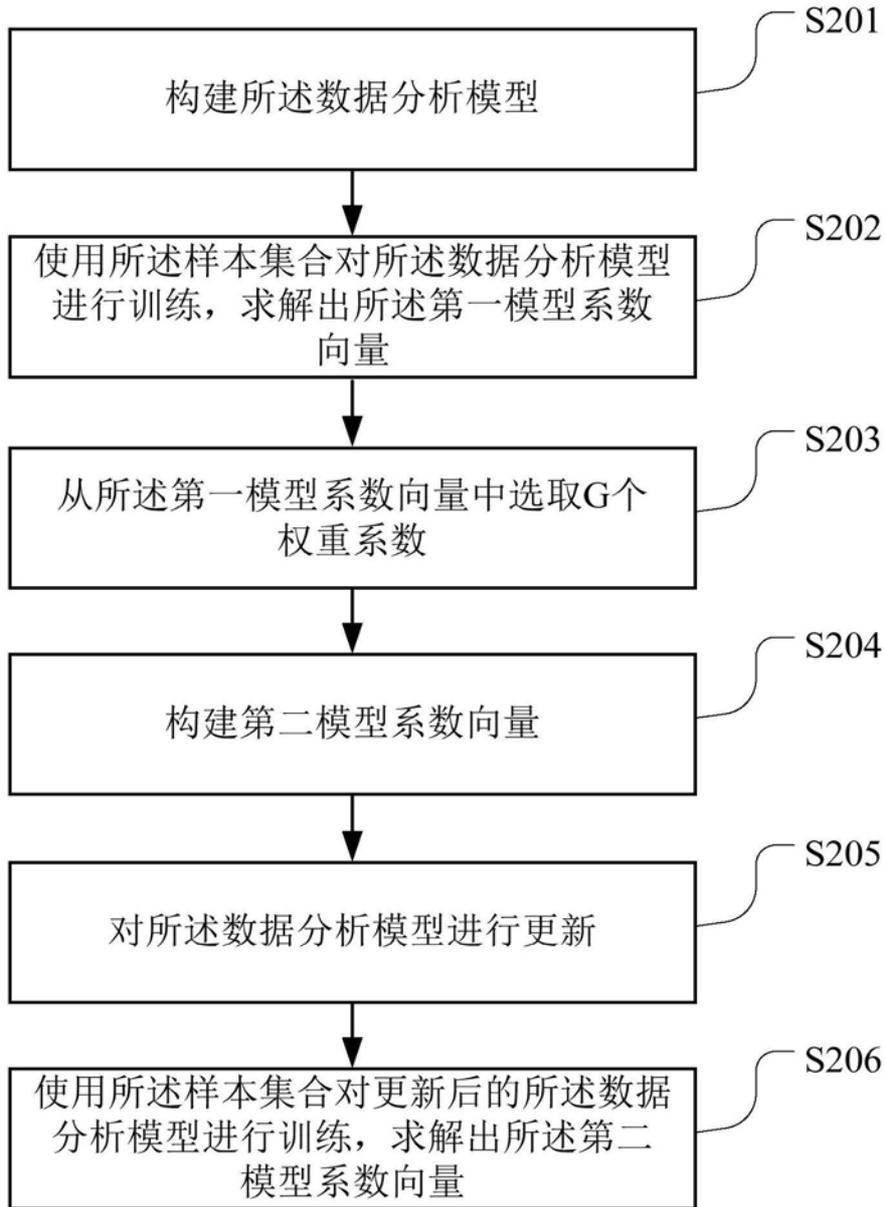


图2

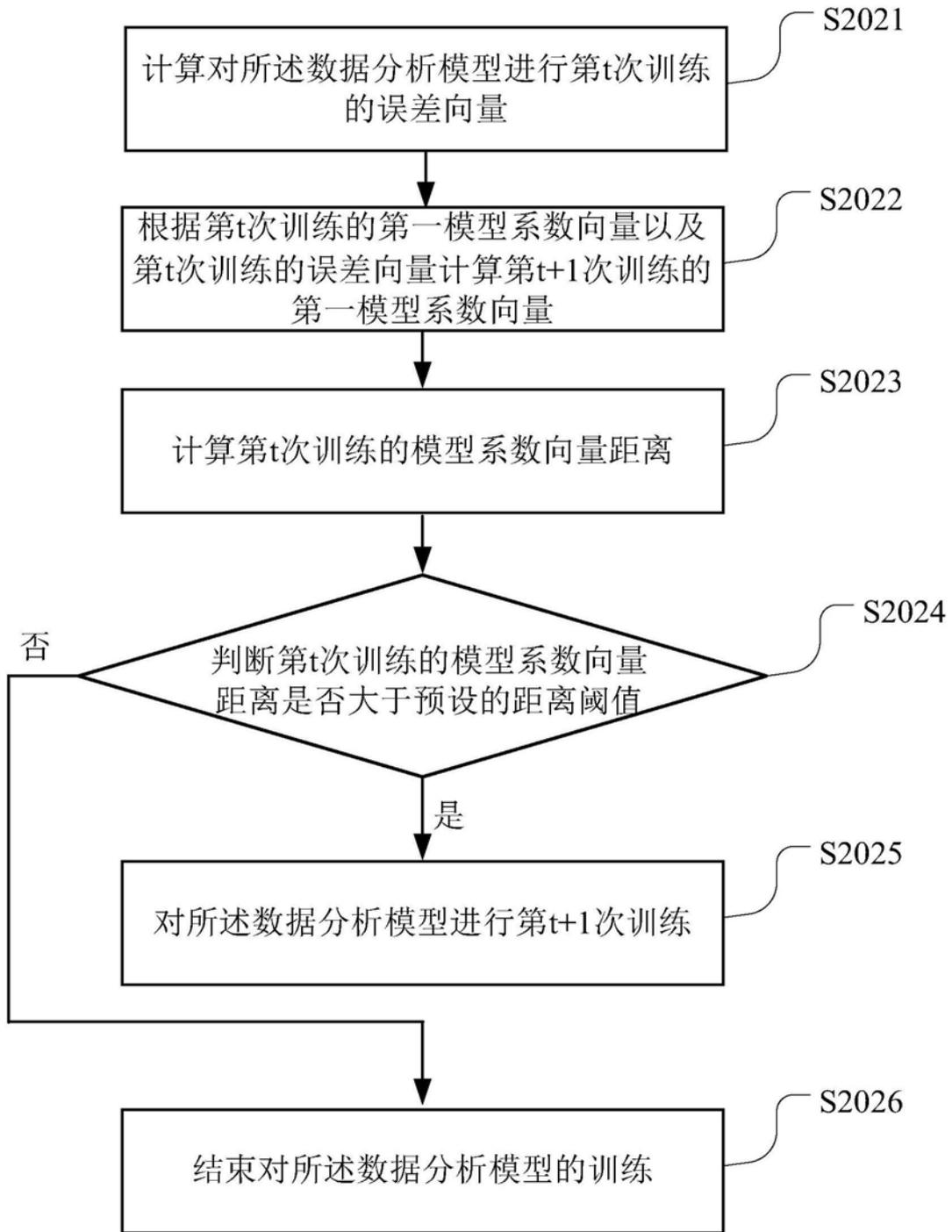


图3

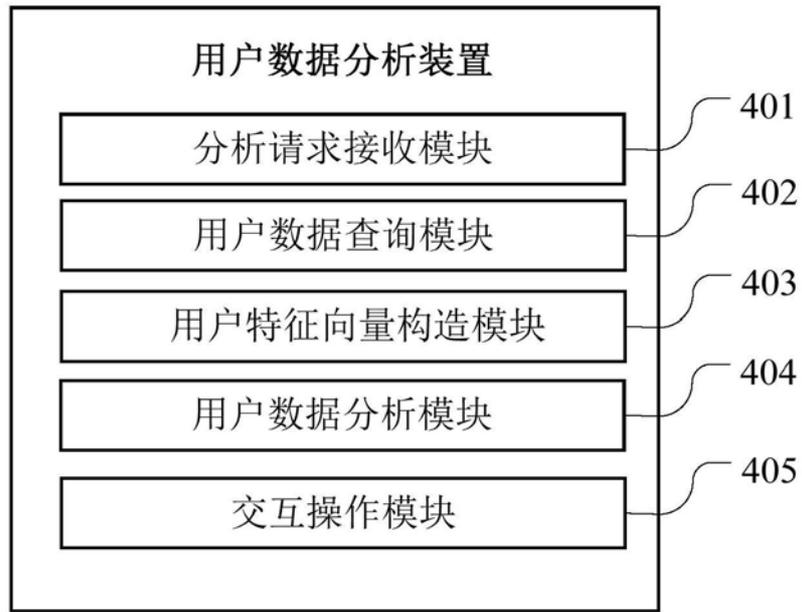


图4

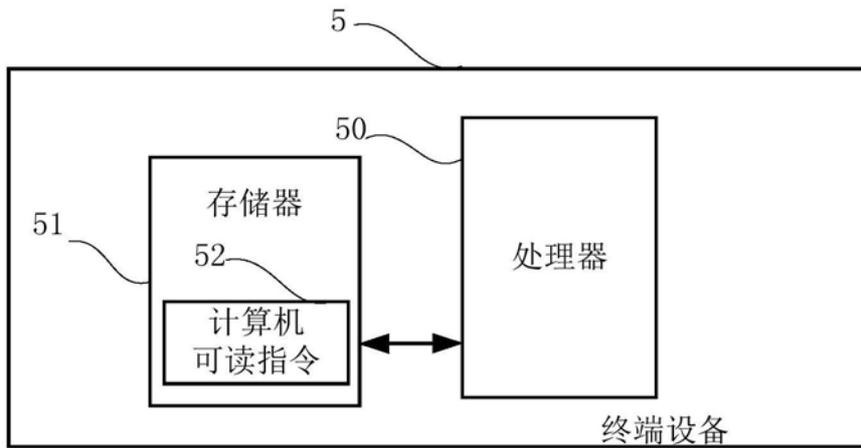


图5