(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112200375 B
(45)授权公告日 2023.08.29

(21)申请号 202011105677.3

(22)申请日 2020.10.15

(65)同一申请的已公布的文献号
申请公布号 CN 112200375 A

(43)申请公布日 2021.01.08

(73)专利权人 中国联合网络通信集团有限公司
地址 100033 北京市西城区金融大街21号

(72)发明人 魏进武 崔羽飞 张第

(74)专利代理机构 北京天昊联合知识产权代理有限公司 11112
专利代理师 彭瑞欣 吴俣

(51)Int.Cl.
G06Q 10/04(2023.01)
G06Q 50/32(2012.01)
G06F 16/28(2019.01)
G06F 16/25(2019.01)
G06F 16/182(2019.01)
G06F 18/214(2023.01)

(56)对比文件
CN 110837931 A,2020.02.25
US 2009248475 A1,2009.10.01
CN 110971460 A,2020.04.07
CN 109558962 A,2019.04.02
CN 107482675 A,2017.12.15
CN 108712279 A,2018.10.26
Zhang di等.Numerical prediction on turbine blade internal tip cooling with pin-fin and dimple/protrusion structures.《NUMERICAL HEAT TRANSFER PART A-APPLICATIONS》.2016,第70卷(第09期),第1021-1040页,全文.

审查员 范琳琳

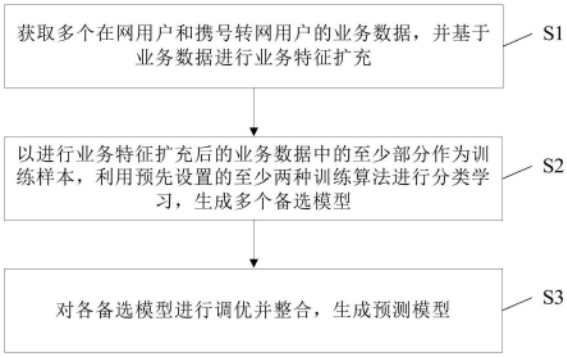
权利要求书2页 说明书6页 附图3页

(54)发明名称

预测模型生成方法、预测模型生成装置和计算机可读介质

(57)摘要

本公开提供了一种预测模型生成方法,包括:获取多个在网用户和携号转网用户的业务数据,并基于业务数据进行业务特征扩充;以进行业务特征扩充后的业务数据中的至少部分作为训练样本,利用预先设置的至少两种训练算法进行分类学习,生成多个备选模型;对各备选模型进行调优并整合,生成预测模型,该预测模型用于根据输入的业务数据输出对应用户的携号转网概率。本公开还提供了一种预测模型生成装置和计算机可读介质。



CN 112200375 B

1.一种预测模型生成方法,其特征在于,包括:

获取多个在网用户和携号转网用户的业务数据,并基于所述业务数据进行业务特征扩充;所述业务数据包括业务使用量和消费数据;

以进行业务特征扩充后的所述业务数据中的至少部分作为训练样本,利用预先设置的至少两种训练算法进行分类学习,生成多个备选模型;

对各所述备选模型进行调优并整合,生成预测模型,所述预测模型用于根据输入的业务数据输出对应用户的携号转网概率;

所述基于所述业务数据进行业务特征扩充包括:根据所述业务使用量和所述消费数据计算得到对应的业务使用趋势;根据所述业务使用趋势计算得到所述业务使用稳定度,并将所述业务使用稳定度添加至所述业务数据中,所述业务数据还包括:语音业务数据、流量业务数据和订阅业务数据中的至少一者,所述业务使用稳定度包括:语音业务使用稳定度、流量业务使用稳定度和订阅业务使用稳定度中的至少一者;

所述根据所述业务使用量和所述消费数据计算得到对应的业务使用趋势包括:采用如下公式:$k_i=\frac{\sum_{m=i-x}^{i+x}(f_m-\overline{f})(l-\overline{l})}{\sum_{m=i-x}^{i+x}(l-\overline{l})^2}$计算所述业务数据在第i月时对应的业务使用趋势$k_i$;其中,$f_m$表示第i月的业务使用量;$\overline{f}=\frac{1}{2x}\sum_{m=i-x}^{i+x}f_m$,表示第i月前后x个月的业务使用量的斜率;l表示第i月的消费数据;$\overline{l}=\frac{1}{2x}\sum_{m=i-x}^{i+x}l$,表示第i月前后x个月的消费数据的斜率;

所述根据所述业务使用趋势计算得到所述业务使用稳定度,包括:采用如下公式:$W=\sum_{n=i-x}^{i+x}M(n)$计算所述业务使用稳定度W;其中,$M(n)=\begin{cases}1 & , k_i<k_{i-1}\\ 0 & , k_i\geq k_{i-1}\end{cases}$,表示所述业务数据在第n月时对应的稳定度系数;

所述对各所述备选模型进行调优并整合,生成预测模型包括:利用网格搜索对各所述备选模型进行调优,并利用堆叠算法对调优后的各所述备选模型进行整合,生成所述预测模型。

2.根据权利要求1所述的预测模型生成方法,其特征在于,所述获取多个在网用户和携号转网用户的业务数据的步骤,包括:

从数据仓库中获取所述业务数据,所述数据仓库包括:数据库、分布式文件系统和蜂巢存储系统;

在所述获取多个在网用户和携号转网用户的业务数据的步骤之后,所述根据所述业务数据进行业务特征扩充的步骤之前,还包括:

利用SparkSQL对所述业务数据进行特征筛选,并将筛选后与所述预测模型相关的数据字段存储至所述分布式文件系统中;

利用SparkSQL对存储至所述分布式文件系统中的所述业务数据进行数据预处理,生成用于进行业务特征扩充的所述业务数据,其中,所述数据预处理包括:数据转换、数据探索、属性规约和数据标准化中的至少一者。

3.根据权利要求1所述的预测模型生成方法，其特征在于，所述以进行业务特征扩充后的所述业务数据中的至少部分作为训练样本，利用预先设置的至少两种训练算法进行分类学习，生成多个备选模型的步骤，包括：

以进行业务特征扩充后的所述业务数据中的部分作为训练样本，利用至少两种训练算法进行分类学习，生成多个待优化模型，其中，所述训练算法包括：逻辑回归算法、决策树算法、随机森林算法和极端梯度提升算法；

以进行业务特征扩充后的所述业务数据中的另一部分作为测试样本，对训练出的全部所述待优化模型进行优化，生成多个所述备选模型。

4.一种预测模型生成装置，包括：

一个或多个处理器；

存储单元，用于存储一个或多个程序；

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-3中任一所述的预测模型生成方法。

5.一种计算机可读介质，其上存储有计算机程序，其中，所述程序被处理器执行时实现如权利要求1-3中任一所述的预测模型生成方法中的步骤。

预测模型生成方法、预测模型生成装置和计算机可读介质

技术领域

[0001] 本公开涉及通信技术领域，特别涉及一种预测模型生成方法、预测模型生成装置和计算机可读介质。

背景技术

[0002] 随着通信技术的发展和广泛应用，越来越多的人们成为电信运营商的用户，电信运营商之间的商业竞争也愈加激烈。目前，各运营商已具备携号转网的基础要求，用户可根据自身的需求使用该业务在不更换号码的前提下进行运营商更换。现阶段暂未落实对可能进行携号转网的用户进行预测评估的具体手段，运营商无法对用户可能携号转网的概率进行预测，由此对后续的业务提供产生影响。

发明内容

[0003] 本公开旨在至少解决现有技术中存在的技术问题之一，提出了一种预测模型生成方法、预测模型生成装置和计算机可读介质。

[0004] 为实现上述目的，第一方面，本公开实施例提供了一种预测模型生成方法，包括：

[0005] 获取多个在网用户和携号转网用户的业务数据，并基于所述业务数据进行业务特征扩充；

[0006] 以进行业务特征扩充后的所述业务数据中的至少部分作为训练样本，利用预先设置的至少两种训练算法进行分类学习，生成多个备选模型；

[0007] 对各所述备选模型进行调优并整合，生成预测模型，所述预测模型用于根据输入的业务数据输出对应用户的携号转网概率。

[0008] 在一些实施例中，所述获取多个在网用户和携号转网用户的业务数据的步骤，包括：

[0009] 从数据仓库中获取所述业务数据，所述数据仓库包括：数据库、分布式文件系统和蜂巢存储系统；

[0010] 在所述获取多个在网用户和携号转网用户的业务数据的步骤之后，所述根据所述业务数据进行业务特征扩充的步骤之前，还包括：

[0011] 利用SparkSQL对所述业务数据进行特征筛选，并将筛选后与所述预测模型相关的数据字段存储至所述分布式文件系统中；

[0012] 利用SparkSQL对存储至所述分布式文件系统中的所述业务数据进行数据预处理，生成用于进行业务特征扩充的所述业务数据，其中，所述数据预处理包括：数据转换、数据探索、属性规约和数据标准化中的至少一者。

[0013] 在一些实施例中，所述基于所述业务数据进行业务特征扩充的步骤，包括：

[0014] 计算所述业务数据对应的业务使用稳定度，并将所述业务使用稳定度添加至所述业务数据中，其中，所述业务数据包括：语音业务数据、流量业务数据和订阅业务数据中的至少一者，所述业务使用稳定度包括：语音业务使用稳定度、流量业务使用稳定度和订阅业

务使用稳定度中的至少一者。

[0015] 在一些实施例中，所述业务数据还包括：业务使用量和消费数据；

[0016] 所述计算所述业务数据对应的业务使用稳定度的步骤，包括：

[0017] 根据所述业务使用量和所述消费数据计算得到对应的业务使用趋势；

[0018] 根据所述业务使用趋势计算得到所述业务使用稳定度。

[0019] 在一些实施例中，所述根据所述业务使用量计算得到对应的业务使用趋势的步骤，包括：

[0020] 采用如下公式：

[0021] $$k_i=\frac{\sum_{m=i-x}^{i+x}(f_m-\overline{f})(l-\overline{l})}{\sum_{m=i-x}^{i+x}(l-\overline{l})^2}$$

[0022] 计算所述业务数据在第i月时对应的业务使用趋势k_i；其中，f_m表示第i月的业务使用量；$\overline{f}=\frac{1}{2x}\sum_{m=i-x}^{i+x}f_m$，表示第i月前后x个月的业务使用量的斜率；l表示第i月的消费数据；$\overline{l}=\frac{1}{2x}\sum_{m=i-x}^{i+x}l$，表示第i月前后x个月的消费数据的斜率。

[0023] 在一些实施例中，所述根据所述业务使用趋势计算得到所述业务使用稳定度的步骤，包括：

[0024] 采用如下公式：

[0025] $$W=\sum_{n=i-x}^{i+x}M(n)$$

[0026] 计算所述业务使用稳定度W；其中，$M(n)=\begin{cases}1 & , \ k_i<k_{i-1}\\ 0 & , \ k_i\geq k_{i-1}\end{cases}$，表示所述业务数据在第n月时对应的稳定度系数。

[0027] 在一些实施例中，所述以进行业务特征扩充后的所述业务数据中的至少部分作为训练样本，利用预先设置的至少两种训练算法进行分类学习，生成多个备选模型的步骤，包括：

[0028] 以进行业务特征扩充后的所述业务数据中的部分作为训练样本，利用至少两种训练算法进行分类学习，生成多个待优化模型，其中，所述训练算法包括：逻辑回归算法、决策树算法、随机森林算法和极端梯度提升算法；

[0029] 以进行业务特征扩充后的所述业务数据中的另一部分作为测试样本，对训练出的全部所述待优化模型进行优化，生成多个所述备选模型。

[0030] 在一些实施例中，所述对各所述备选模型进行调优并整合，生成预测模型的步骤，包括：

[0031] 利用网格搜索对各所述备选模型进行调优，并利用堆叠算法对调优后的各所述备选模型进行整合，生成所述预测模型。

[0032] 第二方面，本公开实施例还提供了一种预测模型生成装置，包括：

[0033] 一个或多个处理器；

[0034] 存储单元，用于存储一个或多个程序；

[0035] 当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如上述实施例中任一所述的预测模型生成方法。

[0036] 第三方面，本公开实施例还提供了一种计算机可读介质，其上存储有计算机程序，其中，所述程序被处理器执行时实现如上述实施例中任一所述的预测模型生成方法中的步骤。

[0037] 本公开具有以下有益效果：

[0038] 本公开实施例提供了一种预测模型生成方法、预测模型生成装置和计算机可读介质，可通过获取不同用户的业务数据，并基于进行业务特征扩充后的该业务数据利用多种算法进行学习训练，生成预测模型，该预测模型用于根据输入的业务数据输出对应用户的携号转网概率，实现更精确地对携号转网用户进行预测和评估。

附图说明

[0039] 图1为本公开实施例提供的一种预测模型生成方法的流程图；

[0040] 图2为本公开实施例中步骤S2的一种具体实施方法流程图；

[0041] 图3为本公开实施例中步骤S1的一种具体实施方法流程图；

[0042] 图4为本公开实施例中步骤S3的一种具体实施方法流程图；

[0043] 图5为本公开实施例中步骤S301的一种具体实施方法流程图。

具体实施方式

[0044] 为使本领域的技术人员更好地理解本公开的技术方案，下面结合附图对本公开提供的预测模型生成方法、预测模型生成装置和计算机可读介质进行详细描述。

[0045] 本公开所提供的预测模型生成方法、预测模型生成装置和计算机可读介质，可用于通过获取不同用户的业务数据，并基于进行业务特征扩充后的该业务数据利用多种算法进行学习训练，生成预测模型，该预测模型用于根据输入的业务数据输出对应用户的携号转网概率，实现更精确地对携号转网用户进行预测和评估。

[0046] 图1为本公开实施例提供的一种预测模型生成方法的流程图。

[0047] 如图1所示，该方法包括：

[0048] 步骤S1、获取多个在网用户和携号转网用户的业务数据，并基于业务数据进行业务特征扩充。

[0049] 其中，获取多个在网用户和携号转网用户的业务数据，即进行数据准备，基于业务数据进行业务特征扩充，即特征工程；业务数据可包括消费数据、语音业务数据、流量业务数据和订阅业务数据等；消费数据可包括套餐内费用、套餐外费用以及各项业务对应的单项费用；语音业务数据可包括主叫通话时间和被叫通话时间等；流量业务数据可包括本地流量、省际漫游流量、国际漫游流量和港澳台漫游流量等；订阅业务数据针对应用程序订阅和其他附加业务(如定制彩铃)等；具体以是否进行携号转网出账作为是否为携号转网用户的依据。

[0050] 在一些实施例中，在进行业务特征扩充之后还包括：对进行业务特征扩充后的业务数据中的各特征进行评分，选取评分较高的特征进行后续算法训练，提高预测的准确性。

[0051] 步骤S2、以进行业务特征扩充后的业务数据中的至少部分作为训练样本，利用预先设置的至少两种训练算法进行分类学习，生成多个备选模型。

[0052] 图2为本公开实施例中步骤S2的一种具体实施方法流程图。如图2所示，步骤S2，以进行业务特征扩充后的业务数据中的至少部分作为训练样本，利用预先设置的至少两种训练算法进行分类学习，生成多个备选模型的步骤，包括：步骤S201和步骤S202。

[0053] 步骤S201、以进行业务特征扩充后的业务数据中的部分作为训练样本，利用至少两种训练算法进行分类学习，生成多个待优化模型。

[0054] 其中，训练算法包括：逻辑回归算法、决策树算法、随机森林算法和极端梯度提升算法(eXtreme Gradient Boosting，简称XGBoost)。

[0055] 步骤S202、以进行业务特征扩充后的业务数据中的另一部分作为测试样本，对训练出的全部待优化模型进行优化，生成多个备选模型。

[0056] 其中，为了交叉验证，利用SparkSQL对业务数据进行分割。在一些实施例中，对业务数据进行7:3的分割，即全部业务数据中的70％用于模型训练，30％用于模型验证。

[0057] 步骤S3、对各备选模型进行调优并整合，生成预测模型。

[0058] 其中，该预测模型用于根据输入的业务数据输出对应用户的携号转网概率。

[0059] 在一些实施例中，在步骤S3中，对各备选模型进行调优并整合，生成预测模型的步骤，包括：利用网格搜索(Grid Search)对各备选模型进行调优，并利用堆叠算法(Stacking)对调优后的各备选模型进行整合，生成预测模型。

[0060] 在一些实施例中，在生成预测模型之后还包括：将该预测模型上传至区块链预测全模型中，此后利用区块链网络的全模型参与和智能合约等特性对用户是否会携号转网进行综合预测。

[0061] 本公开实施例提供了一种预测模型生成方法，该方法可用于通过获取不同用户的业务数据，并基于进行业务特征扩充后的该业务数据利用多种算法进行学习训练，整合多种算法训练而成的备选模型生成预测模型，该预测模型用于根据输入的业务数据输出对应用户的携号转网概率，利用更多维的参数构建预测模型，实现更精确地对携号转网用户进行预测和评估。

[0062] 图3为本公开实施例中步骤S1的一种具体实施方法流程图。如图3所示，在步骤S1中，获取多个在网用户和携号转网用户的业务数据的步骤，具体包括：步骤S101；在步骤S1中，获取多个在网用户和携号转网用户的业务数据的步骤之后，以及根据业务数据进行业务特征扩充的步骤之前，还包括：步骤S102和步骤S103。

[0063] 步骤S101、从数据仓库中获取业务数据。

[0064] 其中，数据仓库包括数据库、分布式文件系统(Hadoop Distributed File System，简称HDFS)和蜂巢存储系统(HIVE)。

[0065] 步骤S102、利用SparkSQL对业务数据进行特征筛选，并将筛选后与预测模型相关的数据字段存储至分布式文件系统中。

[0066] 其中，利用SparkSQL根据预测需求从业务数据中挑选相应的字段，并存储至分布式文件系统中。

[0067] 步骤S103、利用SparkSQL对存储至分布式文件系统中的业务数据进行数据预处理，生成用于进行业务特征扩充的业务数据。

[0068] 其中，数据预处理包括数据转换、数据探索、属性规约和数据标准化中的至少一者；数据转换为将不同类型的数据进行转换，以使其符合后续处理标准；数据探索后根据探索结果，对异常值和缺失值进行处理，即进行缺失值和异常值过滤；属性规约为删除不相关或弱相关的数据，即数据选择过程。

[0069] 图4为本公开实施例中步骤S3的一种具体实施方法流程图。如图4所示，在步骤S3中，基于业务数据进行业务特征扩充的步骤，包括：步骤S301。

[0070] 步骤S301、计算业务数据对应的业务使用稳定度，并将业务使用稳定度添加至业务数据中。

[0071] 其中，业务数据包括语音业务数据、流量业务数据和订阅业务数据中的至少一者，业务使用稳定度包括语音业务使用稳定度、流量业务使用稳定度和订阅业务使用稳定度中的至少一者。

[0072] 在一些实施例中，在步骤S3中，基于业务数据进行业务特征扩充的步骤，还包括：计算每月平均业务费用，并将每月平均业务费用添加至业务数据中。具体地，每月平均业务费用根据对应业务单项费用和对应业务套餐内外使用情况计算得到。

[0073] 图5为本公开实施例中步骤S301的一种具体实施方法流程图。具体地，业务数据还包括业务使用量和消费数据；如图5所示，在步骤S301中，计算业务数据对应的业务使用稳定度的步骤，包括：步骤S3011和步骤S3012。

[0074] 步骤S3011、根据业务使用量和消费数据计算得到对应的业务使用趋势。

[0075] 在一些实施例中，采用如下公式：

[0076] $$k_i=\frac{\sum_{m=i-x}^{i+x}(f_m-\overline{f})(l-\overline{l})}{\sum_{m=i-x}^{i+x}(l-\overline{l})^2}$$

[0077] 计算业务数据在第i月时对应的业务使用趋势k_i；其中，f_m表示第i月的业务使用量；$\overline{f}=\frac{1}{2x}\sum_{m=i-x}^{i+x}f_m$，表示第i月前后x个月的业务使用量的斜率；l表示第i月的消费数据；$\overline{l}=\frac{1}{2x}\sum_{m=i-x}^{i+x}l$，表示第i月前后x个月的消费数据的斜率。

[0078] 步骤S3012、根据业务使用趋势计算得到业务使用稳定度。

[0079] 在一些实施例中，采用如下公式：

[0080] $$W=\sum_{n=i-x}^{i+x}M(n)$$

[0081] 计算业务使用稳定度W；其中，$M(n)=\begin{cases}1 & , k_i<k_{i-1}\\ 0 & , k_i\geq k_{i-1}\end{cases}$，表示业务数据在第n月时对应的稳定度系数。

[0082] 本公开实施例提供了一种预测模型生成方法，该方法可用于通过获取不同用户的业务数据，将业务使用稳定度添加至业务数据中，实现业务特征扩充，基于进行业务特征扩充后的该业务数据利用多种算法进行学习训练，生成预测模型，利用更多维且更合理的参数构建预测模型，提升预测携号转网用户的准确性。

[0083] 本公开实施例还提供了一种预测模型生成装置，包括：

[0084] 一个或多个处理器;存储单元,用于存储一个或多个程序;当该一个或多个程序被该一个或多个处理器执行,使得该一个或多个处理器实现如上述实施例中的任一预测模型生成方法。

[0085] 本公开实施例还提供了一种计算机可读介质,其上存储有计算机程序,其中,该程序被处理器执行时实现如上述实施例中的任一预测模型生成方法中的步骤。

[0086] 可以理解的是,以上实施方式仅仅是为了说明本公开的原理而采用的示例性实施方式,然而本公开并不局限于此。对于本领域内的普通技术人员而言,在不脱离本公开的精神和实质的情况下,可以做出各种变型和改进,这些变型和改进也视为本公开的保护范围。

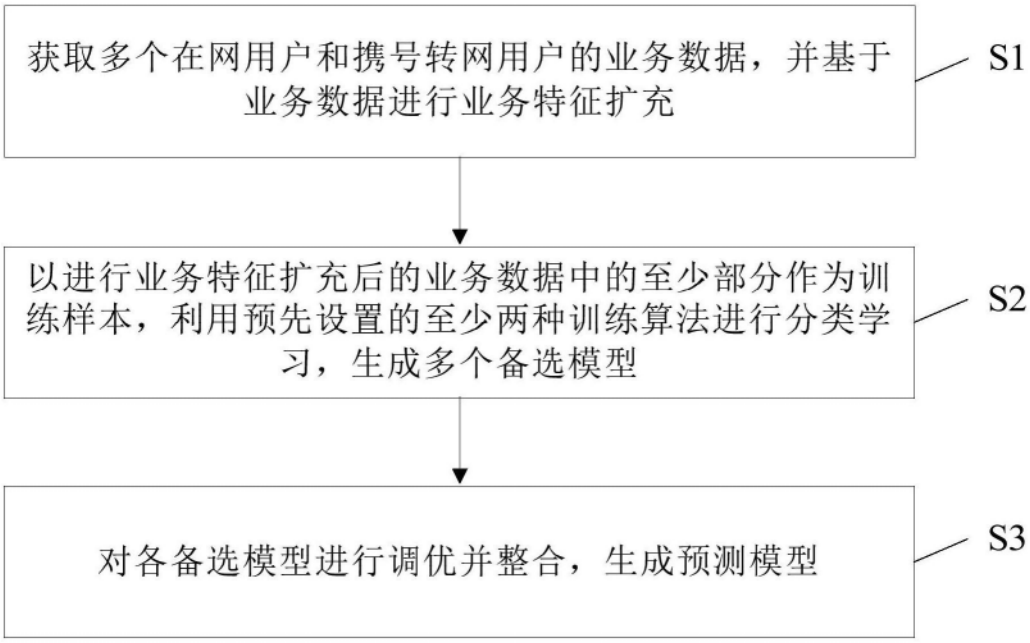

获取多个在网用户和携号转网用户的业务数据，并基于业务数据进行业务特征扩充
S1
以进行业务特征扩充后的业务数据中的至少部分作为训练样本，利用预先设置的至少两种训练算法进行分类学习，生成多个备选模型
S2
对各备选模型进行调优并整合，生成预测模型
S3

图1

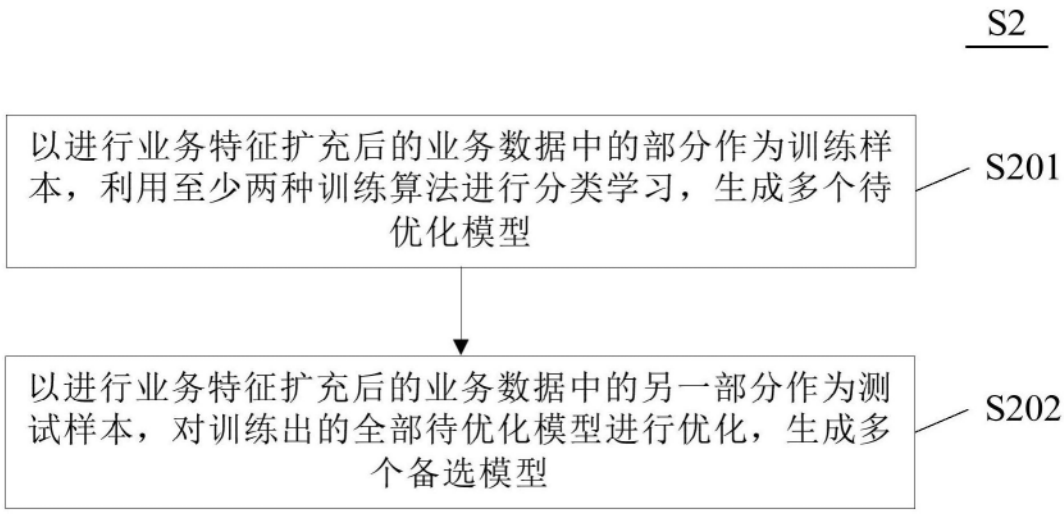

S2
以进行业务特征扩充后的业务数据中的部分作为训练样本，利用至少两种训练算法进行分类学习，生成多个待优化模型
S201
以进行业务特征扩充后的业务数据中的另一部分作为测试样本，对训练出的全部待优化模型进行优化，生成多个备选模型
S202

图2

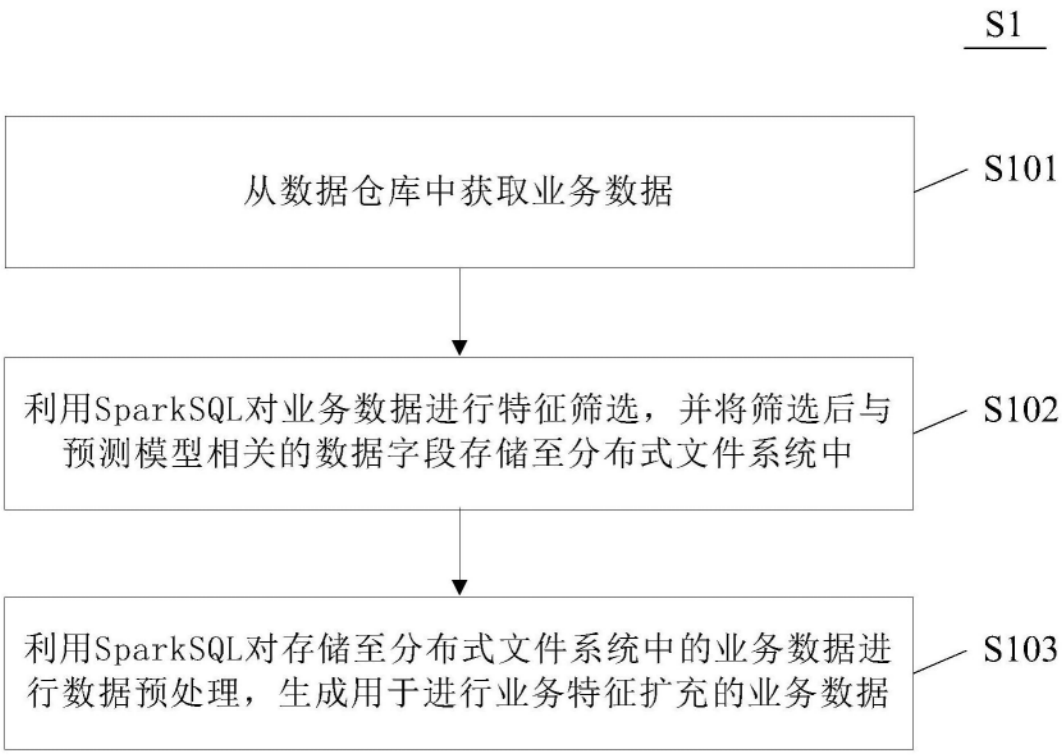

S1
从数据仓库中获取业务数据
S101
利用SparkSQL对业务数据进行特征筛选，并将筛选后与预测模型相关的数据字段存储至分布式文件系统中
S102
利用SparkSQL对存储至分布式文件系统中的业务数据进行数据预处理，生成用于进行业务特征扩充的业务数据
S103

图3

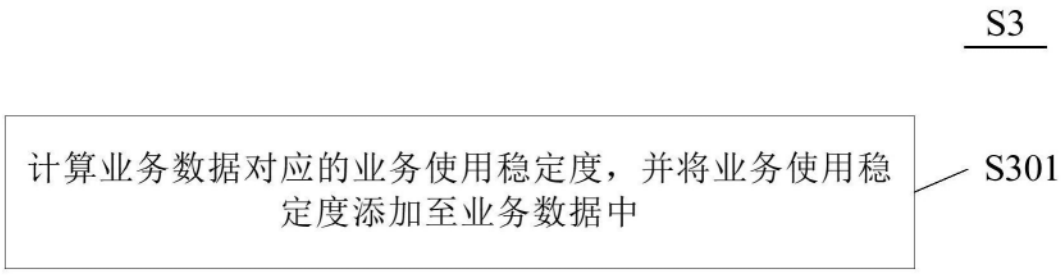

S3
计算业务数据对应的业务使用稳定度，并将业务使用稳定度添加至业务数据中
S301

图4

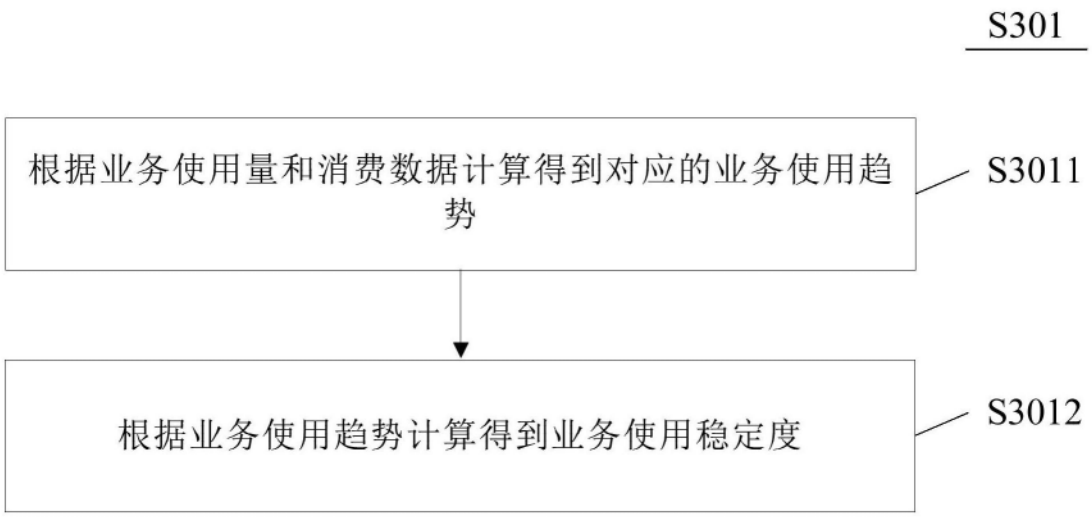

S301
根据业务使用量和消费数据计算得到对应的业务使用趋势
S3011
根据业务使用趋势计算得到业务使用稳定度
S3012

图5