



(12)发明专利申请

(10)申请公布号 CN 108171280 A

(43)申请公布日 2018.06.15

(21)申请号 201810098965.7

(22)申请日 2018.01.31

(71)申请人 国信优易数据有限公司

地址 100070 北京市丰台区南四环西路188号总部广场31号楼

(72)发明人 夏耘海 李燕伟 王甲樑

(74)专利代理机构 北京超凡志成知识产权代理事务所(普通合伙) 11371

代理人 陈剑

(51)Int.Cl.

G06K 9/62(2006.01)

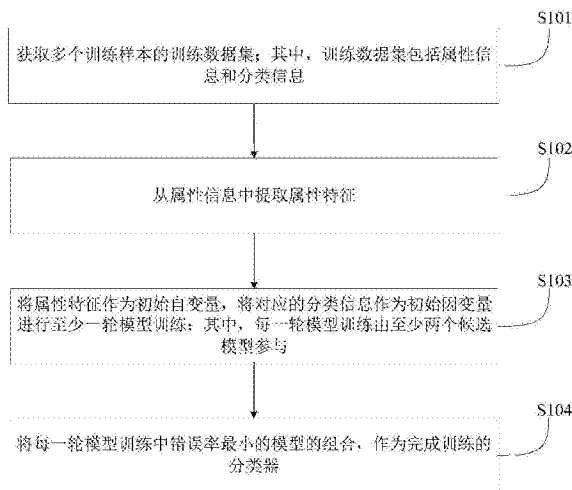
权利要求书2页 说明书11页 附图7页

(54)发明名称

一种分类器构建方法及预测分类的方法

(57)摘要

本发明提供了一种分类器构建方法及预测分类的方法,其中,该分类器构建方法包括:获取多个训练样本的训练数据集;其中,训练数据集包括属性信息和分类信息;从属性信息中提取属性特征;将属性特征作为初始自变量,将对应的分类信息作为初始因变量进行至少一轮模型训练;其中,每一轮模型训练由至少两个候选模型参与;将每一轮模型训练中错误率最小的模型的组合,作为完成训练的分类器。通过本发明提供的分类器构建方法及预测分类的方法,基于至少两个候选模型进行至少一轮模型训练以得到分类器,并能够根据训练好的分类器预测目标样本的分类,避免了采用单一的分类方法所带来的预测精度和预测准确度均较差的问题,预测的精度和准确度均较高。



1. 一种分类器构建方法,其特征在于,包括:

获取多个训练样本的训练数据集;其中,所述训练数据集包括属性信息和分类信息;

从所述属性信息中提取属性特征;

将所述属性特征作为初始自变量,将对应的分类信息作为初始因变量进行至少一轮模型训练;其中,每一轮模型训练由至少两个候选模型参与;

将每一轮模型训练中错误率最小的模型的组合,作为完成训练的分类器。

2. 根据权利要求1所述的方法,其特征在于,每轮训练执行如下操作:

基于本轮训练使用的训练数据确定当前自变量的自变量值和当前因变量的因变量值,对参与本轮训练的至少两个候选模型进行训练;

根据本轮训练的结果,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型;并

确定所述第一候选模型在本轮训练中得到的分类结果错误的出错训练样本;

基于预设权重更新规则,将出错训练样本的权重更新;

根据所述多个训练样本的当前权重,对多个训练样本进行分层抽样处理,得到下一轮训练需要使用的训练数据,进入下一轮训练。

3. 根据权利要求2所述的方法,其特征在于,针对除第一轮训练之外的其他轮训练,在基于本轮训练使用的训练数据确定对应自变量的自变量值和对应因变量的因变量值之前,还包括:

基于上轮训练结束确定的本轮训练需要使用的训练数据所包含训练样本的特征,构建针对本轮训练的新的属性特征;并

将所述新的属性特征确定为当前自变量,将对应的分类信息确定为当前因变量。

4. 根据权利要求2所述的方法,其特征在于,针对除第一轮训练之外的其他轮训练,根据本轮训练的结果,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型,具体包括:

基于所述多个训练样本的属性信息,确定每个训练样本的对应当前自变量的自变量值及当前因变量的因变量值;

将每个训练样本的对应当前自变量值分别输入本轮完成训练的至少两个候选模型,得到各训练样本的分类结果;

根据得到的各训练样本的分类结果以及对应的分类信息,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型。

5. 根据权利要求2-4任一项所述的方法,其特征在于,在确定出错误率最低的第一候选模型之后,还包括:

根据该最低错误率与模型权重的预设数值关系,确定所述第一候选模型的权重;其中,所述预设数值关系满足错误率越小,模型权重越高;

将每一轮模型训练中错误率最小的模型的组合,作为完成训练的分类器,具体包括:

将每一轮模型训练中错误率最小的模型及其对应模型权重的加权组合,作为完成训练的分类器。

6. 根据权利要求2-4任一项所述的方法,其特征在于,根据所述多个训练样本的当前权重,对多个训练样本进行分层抽样处理,得到下一轮训练需要使用的训练数据,具体包括:

从所述多个训练样本中,确定当前权重大于初始权重的部分或全部的训练样本,作为第一训练样本;

根据确定的第一训练样本的数量以及预设数量关系,确定对应数量的当前权重小于初始权重的第二训练样本;并

将所述第一训练样本和所述第二训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据。

7. 根据权利要求6所述的方法,其特征在于,所述将所述第一训练样本和所述第二训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据之前,还包括:

基于所述第一训练样本的分布特征合成预设数量个第三训练样本;

所述将所述第一训练样本和所述第二训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据,包括:

将所述第一训练样本、所述第二训练样本和所述合成的第三训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据。

8. 根据权利要求6所述的方法,其特征在于,在进入下一轮训练之前,还包括:

对确定的下一轮训练需要使用的训练数据进行类别不平衡处理。

9. 根据权利要求2-4任一项所述的方法,其特征在于,在从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型之后,还包括:

确定该最低错误率未达到预设错误率阈值。

10. 一种基于权利要求1至9中任一项训练好的分类器预测分类的方法,其特征在于,包括:

获取目标样本的属性信息;

针对每轮训练得到的错误率最小的模型,基于所述目标样本的属性信息,确定该模型所使用属性特征对应的特征值;

将各错误率最小模型对应的特征值分别输入对应模型得到各错误率最小模型分别对应的分类结果;

基于各错误率最小模型对应的模型权重,对各分类结果进行加权求和,并将得到的和值确定为所述目标样本的分类结果。

## 一种分类器构建方法及预测分类的方法

### 技术领域

[0001] 本发明涉及计算机技术领域,具体而言,涉及一种分类器构建方法及预测分类的方法。

### 背景技术

[0002] 分类算法就是基于分类模型将待检测样本从可选的分类中选取最佳的类别假设,例如,借贷分类模型可以把一个客户对借贷平台的借款意图分类为可能借贷或者不可能借贷。

[0003] 对应于上述分类模型,主要包括模型训练阶段、模型验证阶段和模型应用阶段。其中,上述模型训练阶段,就是要建立模型,使用历史数据集来建立分类模型,模型验证阶段,即是要验证根据历史数据集建立的上述模型,如利用交叉验证方式进行验证,模型应用阶段,即是根据建立的分类模型来预测未知类别的数据。

[0004] 常见的分类算法包括决策树分类、贝叶斯分类,神经网络分类以及逻辑回归分类等分类算法。然而,上述分类算法均采用单一的分类方法,在应用到实际业务数据时,由于算法本身的局限性,不能达到一个较好的预测精度和预测准确度。

### 发明内容

[0005] 有鉴于此,本发明的目的在于提供一种分类器构建方法及预测分类的方法,以提高分类预测的精度和准确度。

[0006] 第一方面,本发明提供了一种分类器构建方法,所述方法包括:

[0007] 获取多个训练样本的训练数据集;其中,所述训练数据集包括属性信息和分类信息;

[0008] 从所述属性信息中提取属性特征;

[0009] 将所述属性特征作为初始自变量,将对应的分类信息作为初始因变量进行至少一轮模型训练;其中,每一轮模型训练由至少两个候选模型参与;

[0010] 将每一轮模型训练中错误率最小的模型的组合,作为完成训练的分类器。

[0011] 结合第一方面,本发明提供了第一方面的第一种可能的实施方式,其中,每轮训练执行如下操作:

[0012] 基于本轮训练使用的训练数据确定当前自变量的自变量值和当前因变量的因变量值,对参与本轮训练的至少两个候选模型进行训练;

[0013] 根据本轮训练的结果,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型;并

[0014] 确定所述第一候选模型在本轮训练中得到的分类结果错误的出错训练样本;

[0015] 基于预设权重更新规则,将出错训练样本的权重更新;

[0016] 根据所述多个训练样本的当前权重,对多个训练样本进行分层抽样处理,得到下一轮训练需要使用的训练数据,进入下一轮训练。

[0017] 结合第一方面的第一种可能的实施方式,本发明提供了第一方面的第二种可能的实施方式,其中,针对除第一轮训练之外的其他轮训练,在基于本轮训练使用的训练数据确定对应自变量的自变量值和对应因变量的因变量值之前,还包括:

[0018] 基于上轮训练结束确定的本轮训练需要使用的训练数据所包含训练样本的特征,构建针对本轮训练的新的属性特征;并

[0019] 将所述新的属性特征确定为当前自变量,将对应的分类信息确定为当前因变量。

[0020] 结合第一方面的第一种可能的实施方式,本发明提供了第一方面的第三种可能的实施方式,其中,针对除第一轮训练之外的其他轮训练,根据本轮训练的结果,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型,具体包括:

[0021] 基于所述多个训练样本的属性信息,确定每个训练样本的对应当前自变量的自变量值及当前因变量的因变量值;

[0022] 将每个训练样本的对应自变量值分别输入本轮完成训练的至少两个候选模型,得到各训练样本的分类结果;

[0023] 根据得到的各训练样本的分类结果以及对应的分类信息,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型。

[0024] 结合第一方面的第一种可能的实施方式至第三种可能的实施方式中任一可能的实施方式,本发明提供了第一方面的第四种可能的实施方式,其中,在确定出错误率最低的第一候选模型之后,还包括:

[0025] 根据该最低错误率与模型权重的预设数值关系,确定所述第一候选模型的权重;其中,所述预设数值关系满足错误率越小,模型权重越高;

[0026] 将每一轮模型训练中错误率最小的模型的组合,作为完成训练的分类器,具体包括:

[0027] 将每一轮模型训练中错误率最小的模型及其对应模型权重的加权组合,作为完成训练的分类器。

[0028] 结合第一方面的第一种可能的实施方式至第三种可能的实施方式中任一可能的实施方式,本发明提供了第一方面的第五种可能的实施方式,其中,根据所述多个训练样本的当前权重,对多个训练样本进行分层抽样处理,得到下一轮训练需要使用的训练数据,具体包括:

[0029] 从所述多个训练样本中,确定当前权重大于初始权重的部分或全部的训练样本,作为第一训练样本;

[0030] 根据确定的第一训练样本的数量以及预设数量关系,确定对应数量的当前权重小于初始权重的第二训练样本;并

[0031] 将所述第一训练样本和所述第二训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据。

[0032] 结合第一方面的第五种可能的实施方式,本发明提供了第一方面的第六种可能的实施方式,其中,所述将所述第一训练样本和所述第二训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据之前,还包括:

[0033] 基于所述第一训练样本的分布特征合成预设数量个第三训练样本;

[0034] 所述将所述第一训练样本和所述第二训练样本对应属性信息以及分类信息作为

下一轮训练需要使用的训练数据,包括:

[0035] 将所述第一训练样本、所述第二训练样本和所述合成的第三训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据。

[0036] 结合第一方面的第五种可能的实施方式,本发明提供了第一方面的第七种可能的实施方式,其中,在进入下一轮训练之前,还包括:

[0037] 对确定的下一轮训练需要使用的训练数据进行类别不平衡处理。

[0038] 结合第一方面的第一种可能的实施方式至第三种可能的实施方式中任一可能的实施方式,本发明提供了第一方面的第八种可能的实施方式,其中,在从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型之后,还包括:

[0039] 确定该最低错误率未达到预设错误率阈值。

[0040] 第二方面,本发明提供了一种基于第一方面、第一方面的第一种可能的实施方式至第八种可能的实施方式中任一可能的实施方式训练好的分类器预测分类的方法,所述方法包括:

[0041] 获取目标样本的属性信息;

[0042] 针对每轮训练得到的错误率最小的模型,基于所述目标样本的属性信息,确定该模型所使用属性特征对应的特征值;

[0043] 将各错误率最小模型对应的特征值分别输入对应模型得到各错误率最小模型分别对应的分类结果;

[0044] 基于各错误率最小模型对应的模型权重,对各分类结果进行加权求和,并将得到的和值确定为所述目标样本的分类结果。

[0045] 本发明提供的分类器构建方法,其首先获取多个训练样本的训练数据集;其中,训练数据集包括属性信息和分类信息;然后从属性信息中提取属性特征;再次将属性特征作为初始自变量,将对应的分类信息作为初始因变量进行至少一轮模型训练;其中,每一轮模型训练由至少两个候选模型参与;最后将每一轮模型训练中错误率最小的模型的组合,作为完成训练的分类器。通过本发明提供的分类器构建方法及预测分类的方法,基于至少两个候选模型进行至少一轮模型训练以得到分类器,能够根据训练好的分类器预测目标样本的分类,避免了采用单一的分类方法所带来的预测精度和预测准确度均较差的问题,预测的精度和准确度均较高。

[0046] 为使本发明的上述目的、特征和优点能更明显易懂,下文特举较佳实施例,并配合所附附图,作详细说明如下。

## 附图说明

[0047] 为了更清楚地说明本发明实施例的技术方案,下面将对实施例中所需要使用的附图作简单地介绍,应当理解,以下附图仅示出了本发明的某些实施例,因此不应被看作是对范围的限定,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,还可以根据这些附图获得其他相关的附图。

[0048] 图1示出了本发明实施例所提供的一种分类器构建方法的流程图;

[0049] 图2示出了本发明实施例所提供的另一种分类器构建方法的流程图;

[0050] 图3示出了本发明实施例所提供的另一种分类器构建方法的流程图;

- [0051] 图4示出了本发明实施例所提供的另一种分类器构建方法的流程图；
- [0052] 图5示出了本发明实施例所提供的另一种分类器构建方法的流程图；
- [0053] 图6示出了本发明实施例所提供的另一种分类器构建方法的流程图；
- [0054] 图7示出了本发明实施例所提供的一种预测分类的方法的流程图；
- [0055] 图8示出了本发明实施例所提供的一种分类器构建装置的结构示意图；
- [0056] 图9示出了本发明实施例所提供的一种计算机设备的结构示意图；
- [0057] 图10示出了本发明实施例所提供的一种预测分类的装置的结构示意图；
- [0058] 图11示出了本发明实施例所提供的一种计算机设备的结构示意图；。

### 具体实施方式

[0059] 为使本发明实施例的目的、技术方案和优点更加清楚，下面将结合本发明实施例中附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。通常在此处附图中描述和示出的本发明实施例的组件可以以各种不同的配置来布置和设计。因此，以下对在附图中提供的本发明的实施例的详细描述并非旨在限制要求保护的本发明的范围，而是仅仅表示本发明的选定实施例。基于本发明的实施例，本领域技术人员在没有做出创造性劳动的前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0060] 考虑到相关分类算法均采用单一的分类方法，在应用到实际业务数据时，由于算法本身的局限性，不能达到一个较好的预测精度和预测准确度。基于此，本发明实施例提供了一种分类器构建方法及预测分类的方法，以提高分类预测的精度和准确度。

[0061] 参见图1，为本发明实施例提供的分类器构建方法的流程图，应用于计算机设备，上述分类器构建方法包括如下步骤：

[0062] S101、获取多个训练样本的训练数据集；其中，训练数据集包括属性信息和分类信息。

[0063] 这里，为了便于对上述训练数据集中的属性信息和分类信息进行理解，现结合借贷预测场景具体阐述本发明实施例获取上述信息的方法。在借贷预测场景中，上述属性信息可以包括但不限于：用户基本信息（如姓名、年龄、职业、身份信息 etc）、购物信息（如购物时间、购物地点等），上述分类信息可以包括是否借贷的记录，以及发生借贷的借贷金额。对于上述属性信息，本发明实施例可以是互联网网站（如天猫、亚马逊等）精确开放的数据接口进行获取，还可以是采用网络爬虫技术，如python（一种面向对象的解释型计算机程序设计语言）实现爬虫的功能，把想要获取的属性信息爬取到本地的计算机设备；对于上述分类信息而言，本发明实施例可以通过获取用户持有的银行卡、信用卡的相关交易信息进行确定，还可以通过用户绑定的网络贷款平台的相关贷款信息进行确定。

[0064] 值得说明的是，上述借贷预测场景仅为一个具体示例，本发明实施例提供的分类器构建方法可以对各种应用场景下的实际业务数据进行预测分类，适用性较强。

[0065] S102、从属性信息中提取属性特征。

[0066] 这里，属性特征指的是对上述属性信息进行处理后的结果。本发明实施例可以对上述属性信息进行过滤、类型转换、衍生等处理，以得到处理后的属性特征。其中，上述过滤处理指的是对属性信息中的缺失信息、重复信息等进行过滤操作，上述类型转换处理可以

是对属性信息进行归一化处理,以把不同来源的数据统一到一个参考系下,这样比较起来才有意义,上述衍生处理指的是根据属性信息进行统计分析得到的额外属性信息,如对于包括购物信息的属性信息而言,可以通过对购物信息的衍生,得到某用户的平均购物次数、最多花费多少金额、购物的价格区间等相关统计信息。本发明实施例可以基于实际获取的属性信息,自适应的进行特征选取。

[0067] S103、将属性特征作为初始自变量,将对应的分类信息作为初始因变量进行至少一轮模型训练;其中,每一轮模型训练由至少两个候选模型参与;

[0068] S104、将每一轮模型训练中错误率最小的模型的组合,作为完成训练的分类器。

[0069] 本发明实施例对于每一轮模型训练而言,均可以至少有两个候选模型参与,其中,上述候选模型可以是神经网络模型、SVM(Support Vector Machine,支持向量机)模型、logistic(回归)模型中的任意组合,还可以是其他分类模型的任意组合。

[0070] 值得说明的是,本发明实施例可以采用如下方式确定分类器是否达到收敛。第一种方式,本发明实施例可以采用分类器的训练轮数是否达到预设训练轮数(如3轮)的判断方式,如果训练轮数达到预设阈值,则确定上述分类器的输出已达到收敛,若训练次数未达到预设阈值,则确定未达到收敛。第二种方式,本发明实施例还可以采用分类器的输出分类结果和实际分类信息之间的输出误差(如误差0.0001)是否小于预设误差的判断方式,如果输出误差小于预设误差,则确定上述分类器的输出已达到收敛,若输出误差小于或等于预设误差,则确定未达到收敛。不管是上述哪一种判断方式,在确定达到收敛后,将所有轮模型训练中错误率最小的模型的组合作为上述分类器即可。

[0071] 针对上述每轮训练而言,参见图2,本发明实施例提供的分类器构建方法还包括如下内容:

[0072] S201、基于本轮训练使用的训练数据确定当前自变量的自变量值和当前因变量的因变量值,对参与本轮训练的至少两个候选模型进行训练;

[0073] S202、根据本轮训练的结果,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型;

[0074] S203、确定第一候选模型在本轮训练中得到的分类结果错误的出错训练样本;

[0075] S204、基于预设权重更新规则,将出错训练样本的权重更新;

[0076] S205、根据多个训练样本的当前权重,对多个训练样本进行分层抽样处理,得到下一轮训练需要使用的训练数据,进入下一轮训练。

[0077] 这里,本发明实施例中,第一轮训练使用的是初始自变量的自变量值和初始因变量的因变量值,对上述参与第一轮训练的至少两个候选模型进行训练,并基于对全部训练样本的分类结果,从上述至少两个候选模型中确定错误率最低的第一候选模型,以根据该第一候选模型确定的分类结果错误的出错训练样本进行权重更新。与第一轮训练不同的是,对于后续的其他轮,使用的是前一轮确定的训练数据对上述参与至少两个候选模型进行训练,与第一轮训练相同的是,均是基于对全部训练样本的分类结果,从上述至少两个候选模型中确定错误率最低的第一候选模型。

[0078] 其中,为了凸显出错训练样本,本发明实施例中的预设权重更新规则指的是将出错训练样本的权重调高,对应的将正确训练样本的权重调低。

[0079] 在具体实施过程中,为了兼顾分类器的预测准确率和效率,本发明实施例中,每轮



训练使用的候选模型可以相同也可以不同,对于相同的情况不再赘述,对于不同的情况,特举例说明:如上一轮中错误率超过阈值的候选模型,下一轮就不再使用了。

[0080] 本发明实施例中,针对除第一轮训练之外的其他轮训练,参与本轮训练的属性特征随着训练数据的变化而变化,具体的,参见图3,本发明实施例基于下述步骤进行特征更新:

[0081] S301、基于上轮训练结束确定的本轮训练需要使用的训练数据所包含训练样本的特征,构建针对本轮训练的新的属性特征;

[0082] S302、将新的属性特征确定为当前自变量,将对应的分类信息确定为当前因变量。

[0083] 这里,本轮训练均要基于上轮训练确定的本轮训练需要使用的训练数据所包含训练样本的特征,构建新的属性特征,并分别将新的属性特征以及对应的分类信息作为当前自变量和当前因变量,以基于本轮训练使用的训练数据确定当前自变量的自变量值和当前因变量的因变量值,对参与本轮训练的至少两个候选模型进行训练。

[0084] 本发明实施例中,针对除第一轮训练之外的其他轮训练,根据本轮训练的结果,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型,参见图4,具体包括如下步骤:

[0085] S401、基于多个训练样本的属性信息,确定每个训练样本的对应当前自变量的自变量值及当前因变量的因变量值;

[0086] S402、将每个训练样本的对应自变量值分别输入本轮完成训练的至少两个候选模型,得到各训练样本的分类结果;

[0087] S403、根据得到的各训练样本的分类结果以及对应的分类信息,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型。

[0088] 这里,本发明实施例将确定的每个训练样本的对应当前自变量的自变量值及当前因变量的因变量值分别输入本轮完成训练的至少两个候选模型,得到各训练样本的分类结果,并根据各训练样本的分类结果以及对应的分类信息的比较结果,从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型。

[0089] 值得说明的是,本发明实施例在确定错误率最低的第一候选模型后,将根据预设错误率阈值对最低错误率判断,只有在最低错误率未达到预设错误率阈值时,才进入下一轮训练,若最低错误率达到预设错误率阈值(如0.5),则训练到本轮结束,不再继续。

[0090] 另外,本发明实施例在进入下一轮训练之前,还包括:对确定的下一轮训练需要使用的训练数据进行类别不平衡处理。本发明实施例可以采用上采样方法进行不平衡处理,还可以采用下采样方法进行不平衡处理,还可以采用SMOTE(Synthetic Minority Over-sampling Technique)方法进行不平衡处理。其中,上述上采样方法指的是对比例过低的样本(也即,出错训练样本)重复抽样,以使这类样本的特征被模型学习到;上述下采样方法指的是对比例过高的样本(也即,正确训练样本)减少抽样次数,以防止模型过度学习这类型样本的特征;上述SMOTE方法指的是对少数类样本(也即,出错训练样本)进行分析并根据少数类样本人工合成新样本添加到训练数据中,从而避免过拟合问题。

[0091] 考虑到SMOTE方法的优良特性,本发明实施例优选的根据SMOTE方法进行不平衡处理。也即,本发明实施例可以基于第一训练样本的分布特征合成预设数量个第三训练样本,将合成的第三训练样本添加至训练数据中进行训练,以具体实现类别不平衡处理。上述第

三训练样本的预设数量可以根据第一训练样本的分布特征来确定。

[0092] 本发明实施例中的分类器是依赖于每一轮模型训练中错误率最小的模型的,参见图5,基于上述所有轮错误率最小的模型确定分类器,具体通过如下步骤实现:

[0093] S501、根据该最低错误率与模型权重的预设数值关系,确定第一候选模型的权重;其中,预设数值关系满足错误率越小,模型权重越高;

[0094] S502、将每一轮模型训练中错误率最小的模型及其对应模型权重的加权组合,作为完成训练的分类器。

[0095] 这里,预设数值关系可以按照如下公式确定: $\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m}$ ,其中, $\alpha_m$ 表

示第m轮训练确定的错误率最小的候选模型的模型权重, $e_m$ 表示第m轮训练确定的最低错误率。

[0096] 另外,本发明实施例还将每一轮模型训练中错误率最小的模型及基于上述公式确定的模型权重的加权组合,确定完成训练的分类器。

[0097] 另外,本发明实施例中,除第一轮训练之外的其他轮训练,采用的均是根据前一轮的分类结果进行分层抽样处理所确定的训练数据。根据第一轮的分类结果确定下一轮训练需要使用的训练数据,参见图6,具体通过如下步骤实现:

[0098] S601、从多个训练样本中,确定当前权重大于初始权重的部分或全部的训练样本,作为第一训练样本;

[0099] S602、根据确定的第一训练样本的数量以及预设数量关系,确定对应数量的当前权重小于初始权重的第二训练样本;

[0100] S603、将第一训练样本和第二训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据。

[0101] 这里,在对训练样本进行权重更新之后,首先可以从所有训练样本中确定当前权重大于初始权重的部分或全部的训练样本,作为第一训练样本,该第一训练样本对应于出错训练样本,然后根据上述出错训练样本的数量以及出错训练样本与正确训练样本之间的预设数量关系(如出错训练样本的个数等于正确训练样本的个数),确定对应数量的当前权重小于初始权重的第二训练样本,该第二训练样本对应于正确训练样本,基于上述出错训练样本和正确训练样本对应的属性信息以及分类信息作为下一轮训练需要使用的训练数据。

[0102] 同理,根据下一轮的分类结果确定该下一轮的下一轮训练需要使用的训练数据与上述根据第一轮的分类结果确定下一轮训练需要使用的训练数据的具体实现方法类似,以此类推,在此不做赘述。

[0103] 基于上述实施例训练得到的分类器,本发明实施例还提供了一种预测分类的方法,如图7所示,为本发明实施例提供的预测分类的方法的流程图,应用于计算机设备,上述预测分类的方法包括如下步骤:

[0104] S701、获取目标样本的属性信息;

[0105] S702、针对每轮训练得到的错误率最小的模型,基于目标样本的属性信息,确定该模型所使用属性特征对应的特征值;

[0106] S703、将各错误率最小模型对应的特征值分别输入对应模型得到各错误率最小模型分别对应的分类结果；

[0107] S704、基于各错误率最小模型对应的模型权重，对各分类结果进行加权求和，并将得到的和值确定为目标样本的分类结果。

[0108] 这里，首先对获取的目标样本的属性信息进行特征提取，针对每轮训练得到的错误率最小的模型，基于提取出的目标样本的属性信息，确定该模型所使用属性特征对应的特征值，最后将各错误率最小模型对应的特征值分别输入对应模型得到各错误率最小模型分别对应的分类结果，以基于各错误率最小模型对应的模型权重，对各分类结果进行加权求和，并将得到的和值确定为目标样本的分类结果。可见，采用预先训练好的分类器可以快速高效的为目标样本进行分类预测，且预测的精准度较高，自动化程度也较高。

[0109] 基于同一发明构思，本发明实施例中还提供了与分类器构建方法对应的分类器构建装置，由于本发明实施例中的装置解决问题的原理与本发明实施例上述分类器构建方法相似，因此装置的实施可以参见方法的实施，重复之处不再赘述。如图8所示，为本发明实施例所提供的分类器构建装置的结构示意图，该分类器构建装置包括：

[0110] 训练数据获取模块11，用于获取多个训练样本的训练数据集；其中，训练数据集包括属性信息和分类信息；

[0111] 属性特征提取模块12，用于从属性信息中提取属性特征；

[0112] 模型训练模块13，用于将属性特征作为初始自变量，将对应的分类信息作为初始因变量进行至少一轮模型训练；其中，每一轮模型训练由至少两个候选模型参与；

[0113] 分类器构建模块14，用于将每一轮模型训练中错误率最小的模型的组合，作为完成训练的分类器。

[0114] 在具体实施中，上述模型训练模块13，具体用于针对每轮训练执行如下操作：

[0115] 基于本轮训练使用的训练数据确定当前自变量的自变量值和当前因变量的因变量值，对参与本轮训练的至少两个候选模型进行训练；

[0116] 根据本轮训练的结果，从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型；并

[0117] 确定第一候选模型在本轮训练中得到的分类结果错误的出错训练样本；

[0118] 基于预设权重更新规则，将出错训练样本的权重更新；

[0119] 根据多个训练样本的当前权重，对多个训练样本进行分层抽样处理，得到下一轮训练需要使用的训练数据，进入下一轮训练。

[0120] 上述分类器构建装置还包括：

[0121] 属性特征构建模块15，用于基于上轮训练结束确定的本轮训练需要使用的训练数据所包含训练样本的特征，构建针对本轮训练的新的属性特征；并将新的属性特征确定为当前自变量，将对应的分类信息确定为当前因变量。

[0122] 在一种实施方式中，上述模型训练模块13，具体用于基于多个训练样本的属性信息，确定每个训练样本的对应当前自变量的自变量值及当前因变量的因变量值；将每个训练样本的对应当前自变量值分别输入本轮完成训练的至少两个候选模型，得到各训练样本的分类结果；根据得到的各训练样本的分类结果以及对应的分类信息，从参与本轮训练的至少两个候选模型中确定错误率最低的第一候选模型。

[0123] 上述分类器构建装置还包括：

[0124] 模型权重确定模块16,用于根据该最低错误率与模型权重的预设数值关系,确定第一候选模型的权重;其中,预设数值关系满足错误率越小,模型权重越高;

[0125] 上述分类器构建模块14,具体用于将每一轮模型训练中错误率最小的模型及其对应模型权重的加权组合,作为完成训练的分类器。

[0126] 在另一种实施方式中,上述模型训练模块13,具体用于从多个训练样本中,确定当前权重大于初始权重的部分或全部的训练样本,作为第一训练样本;根据确定的第一训练样本的数量以及预设数量关系,确定对应数量的当前权重小于初始权重的第二训练样本;并将第一训练样本和第二训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据。

[0127] 上述分类器构建装置还包括：

[0128] 训练样本合成模块17,用于基于第一训练样本的分布特征合成预设数量个第三训练样本;

[0129] 上述模型训练模块13,具体用于将第一训练样本、第二训练样本和合成的第三训练样本对应属性信息以及分类信息作为下一轮训练需要使用的训练数据。

[0130] 上述分类器构建装置还包括：

[0131] 不平衡处理模块,用于对确定的下一轮训练需要使用的训练数据进行类别不平衡处理。

[0132] 阈值确定模块,用于确定最低错误率未达到预设错误率阈值。

[0133] 对应于图1至图6中的分类器构建方法,本发明实施例还提供了一种计算机设备21,如图9所示,该设备包括存储器21、处理器22及存储在该存储器21上并可在该处理器22上运行的计算机程序,其中,上述处理器22执行上述计算机程序时实现上述分类器构建方法的步骤。

[0134] 具体地,上述存储器21和处理器22能够为通用的存储器21和处理器22,这里不做具体限定,当处理器22运行存储器21存储的计算机程序时,能够执行上述分类器构建方法,从而解决单一的分类方法所带来的预测精度和预测准确度均较差的问题,用以提高预测精度和预测准确度。

[0135] 对应于图1至图6中的分类器构建方法,本发明实施例还提供了一种计算机可读存储介质,该计算机可读存储介质上存储有计算机程序,该计算机程序被处理器22运行时执行上述分类器构建方法的步骤。

[0136] 具体地,该存储介质能够为通用的存储介质,如移动磁盘、硬盘等,该存储介质上的计算机程序被运行时,能够执行上述分类器构建方法,从而解决单一的分类方法所带来的预测精度和预测准确度均较差的问题,用以提高预测精度和预测准确度。

[0137] 基于同一发明构思,本发明实施例中还提供了与预测分类的方法对应的预测分类的装置,由于本发明实施例中的装置解决问题的原理与本发明实施例上述预测分类的方法相似,因此装置的实施可以参见方法的实施,重复之处不再赘述。如图10所示,为本发明实施例所提供的预测分类的装置示意图,该预测分类的装置包括：

[0138] 属性信息获取模块31,用于获取目标样本的属性信息;

[0139] 特征值确定模块32,用于针对每轮训练得到的错误率最小的模型,基于目标样本

的属性信息,确定该模型所使用属性特征对应的特征值;

[0140] 第一分类结果确定模块33,用于将各错误率最小模型对应的特征值分别输入对应模型得到各错误率最小模型分别对应的分类结果;

[0141] 第二分类结果确定模块34,用于基于各错误率最小模型对应的模型权重,对各分类结果进行加权求和,并将得到的和值确定为目标样本的分类结果。

[0142] 对应于图7中的预测分类的方法,本发明实施例还提供了一种计算机设备40,如图11所示,该设备包括存储器41、处理器42及存储在该存储器41上并可在该处理器42上运行的计算机程序,其中,上述处理器42执行上述计算机程序时实现上述预测分类的方法的步骤。

[0143] 具体地,上述存储器41和处理器42能够为通用的存储器41和处理器42,这里不做具体限定,当处理器42运行存储器41存储的计算机程序时,能够执行上述预测分类的方法,从而解决单一的分类方法所带来的预测精度和预测准确度均较差的问题,用以提高预测精度和预测准确度。

[0144] 对应于图7中的预测分类的方法,本发明实施例还提供了一种计算机可读存储介质,该计算机可读存储介质上存储有计算机程序,该计算机程序被处理器42运行时执行上述预测分类的方法的步骤。

[0145] 具体地,该存储介质能够为通用的存储介质,如移动磁盘、硬盘等,该存储介质上的计算机程序被运行时,能够执行上述预测分类的方法,从而解决单一的分类方法所带来的预测精度和预测准确度均较差的问题,用以提高预测精度和预测准确度。

[0146] 在本发明所提供的实施例中,应该理解到,所揭露装置和方法,可以通过其它的方式实现。以上所描述的装置实施例仅仅是示意性的,例如,所述单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,又例如,多个单元或组件可以结合或者可以集成到另一个系统,或一些特征可以忽略,或不执行。另一点,所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些通信接口,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0147] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0148] 另外,在本发明提供的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。

[0149] 所述功能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,可以存储在一个计算机可读取存储介质中。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)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程序代码的介质。

[0150] 应注意到:相似的标号和字母在下面的附图中表示类似项,因此,一旦某一项在一

个附图中被定义,则在随后的附图中不需要对其进行进一步定义和解释,此外,术语“第一”、“第二”、“第三”等仅用于区分描述,而不能理解为指示或暗示相对重要性。

[0151] 最后应说明的是:以上所述实施例,仅为本发明的具体实施方式,用以说明本发明的技术方案,而非对其限制,本发明的保护范围并不局限于此,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:任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,其依然可以对前述实施例所记载的技术方案进行修改或可轻易想到变化,或者对其中部分技术特征进行等同替换;而这些修改、变化或者替换,并不使相应技术方案的本质脱离本发明实施例技术方案的精神和范围。都应涵盖在本发明的保护范围之内。因此,本发明的保护范围应所述以权利要求的保护范围为准。

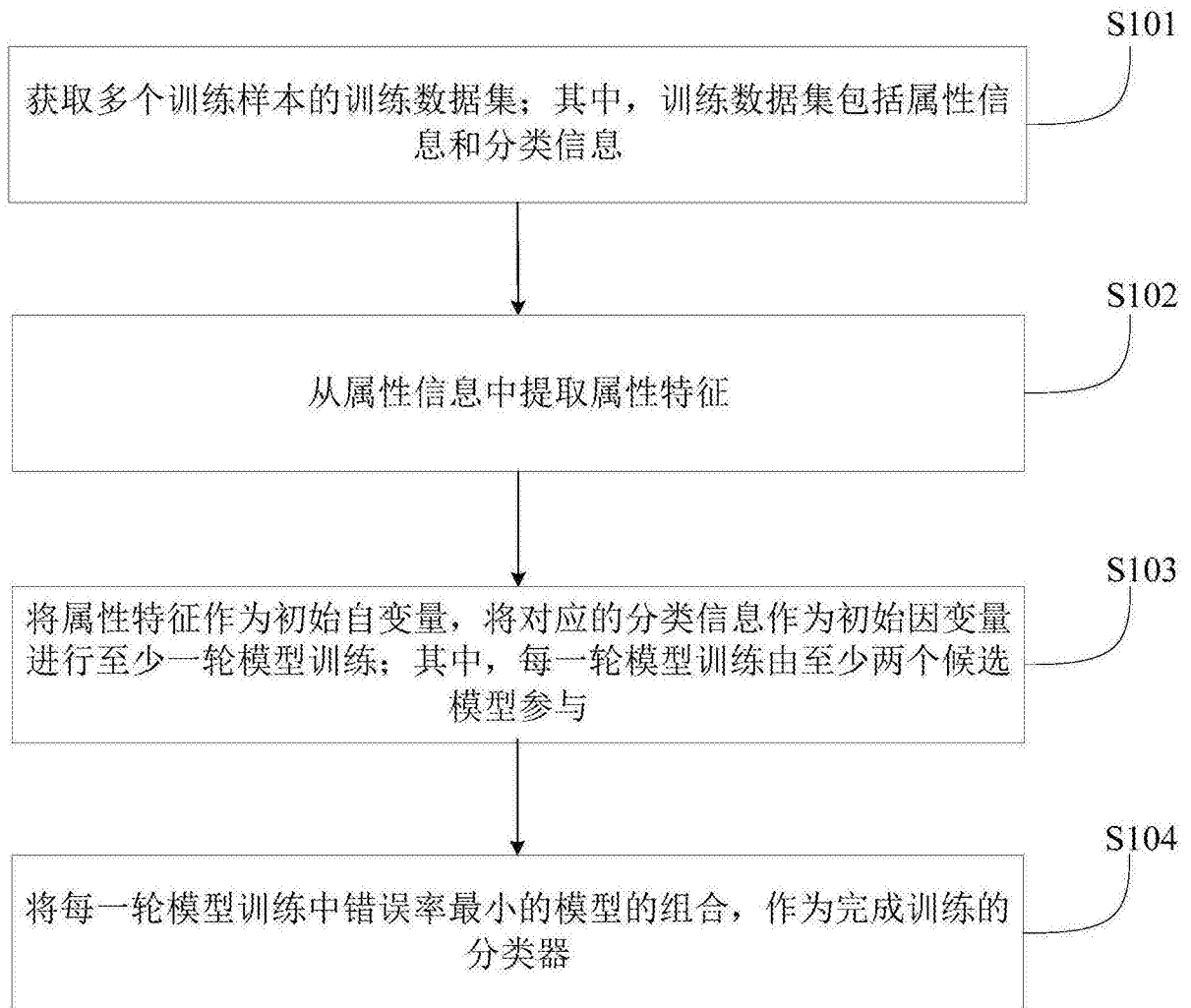


图1

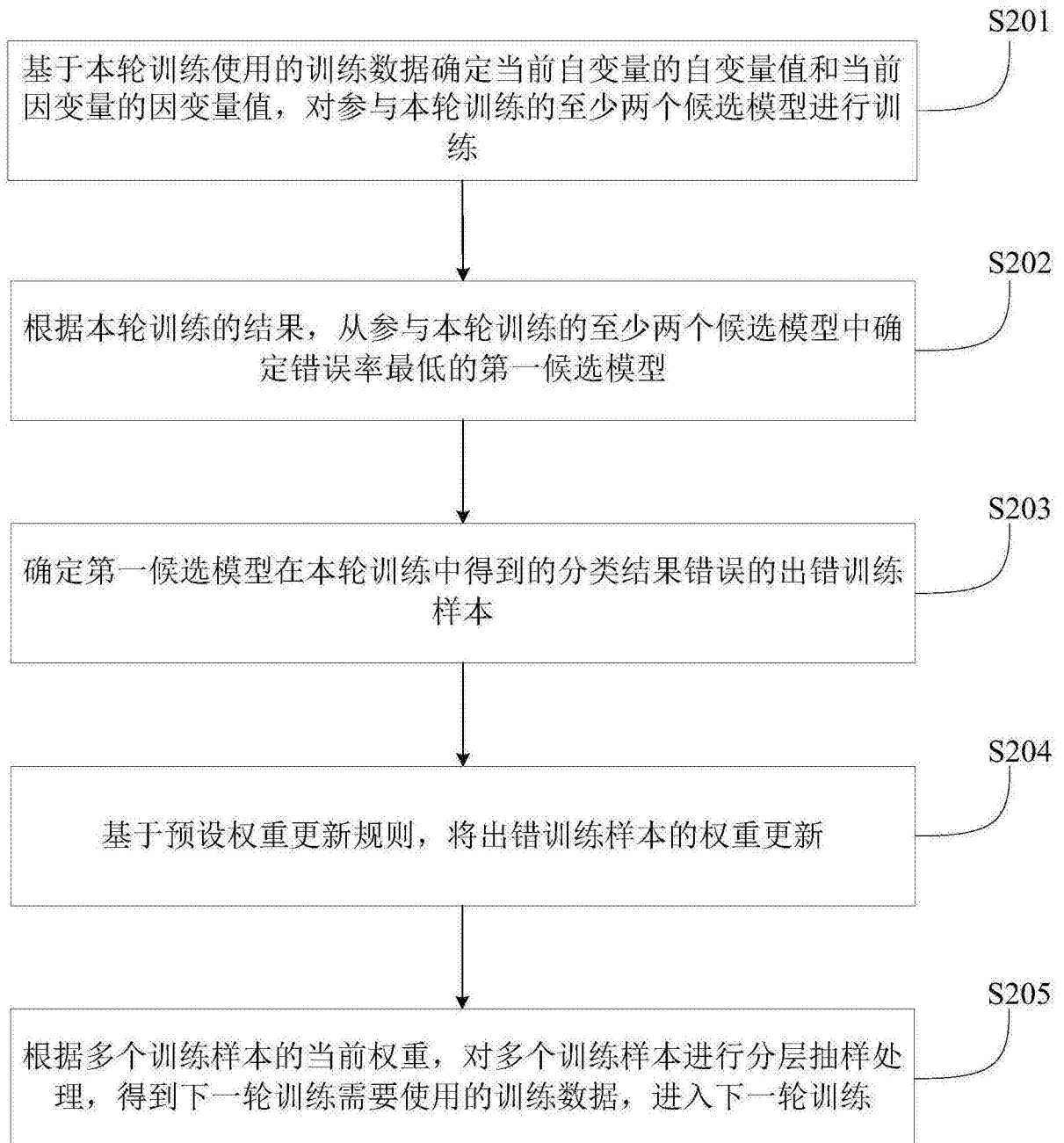


图2



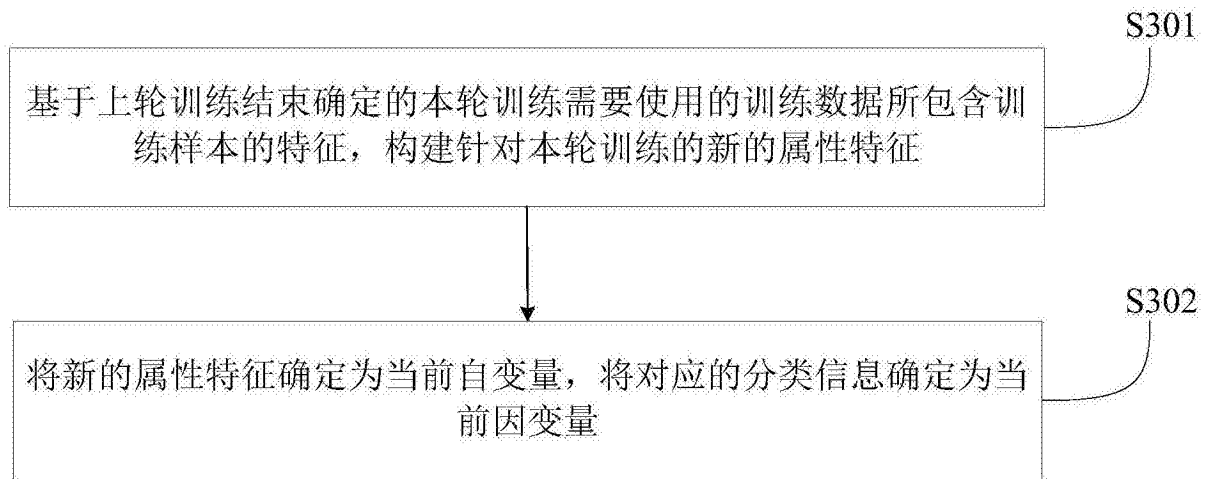


图3

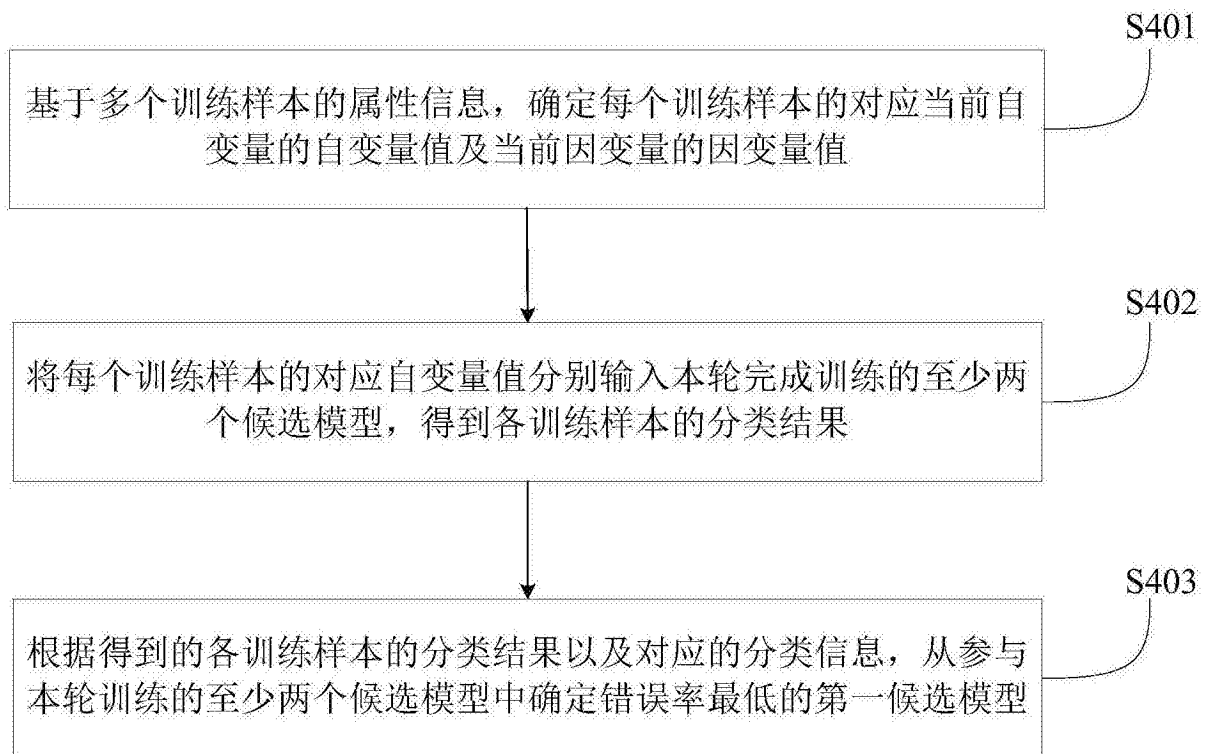


图4

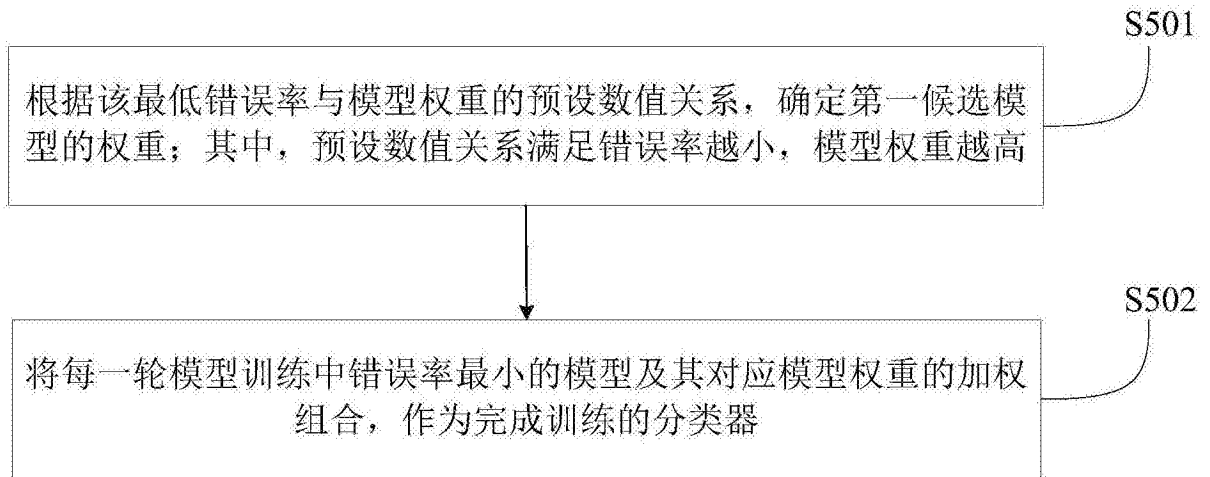


图5

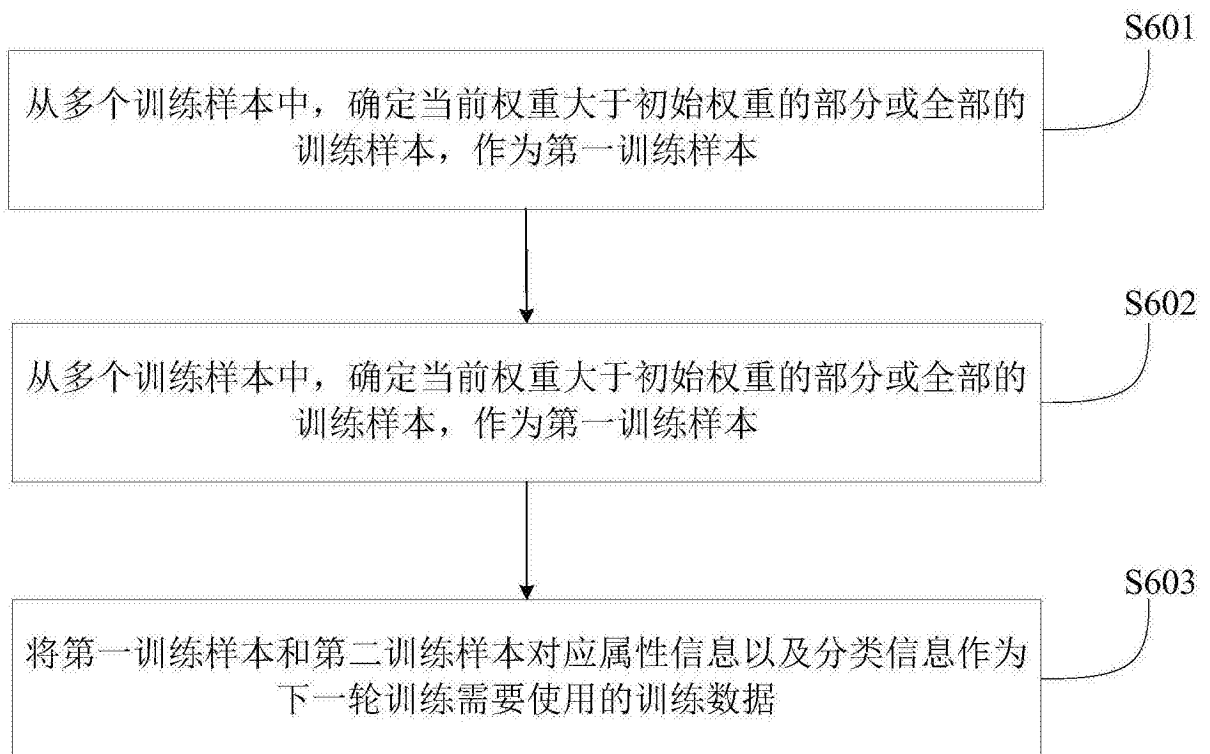


图6

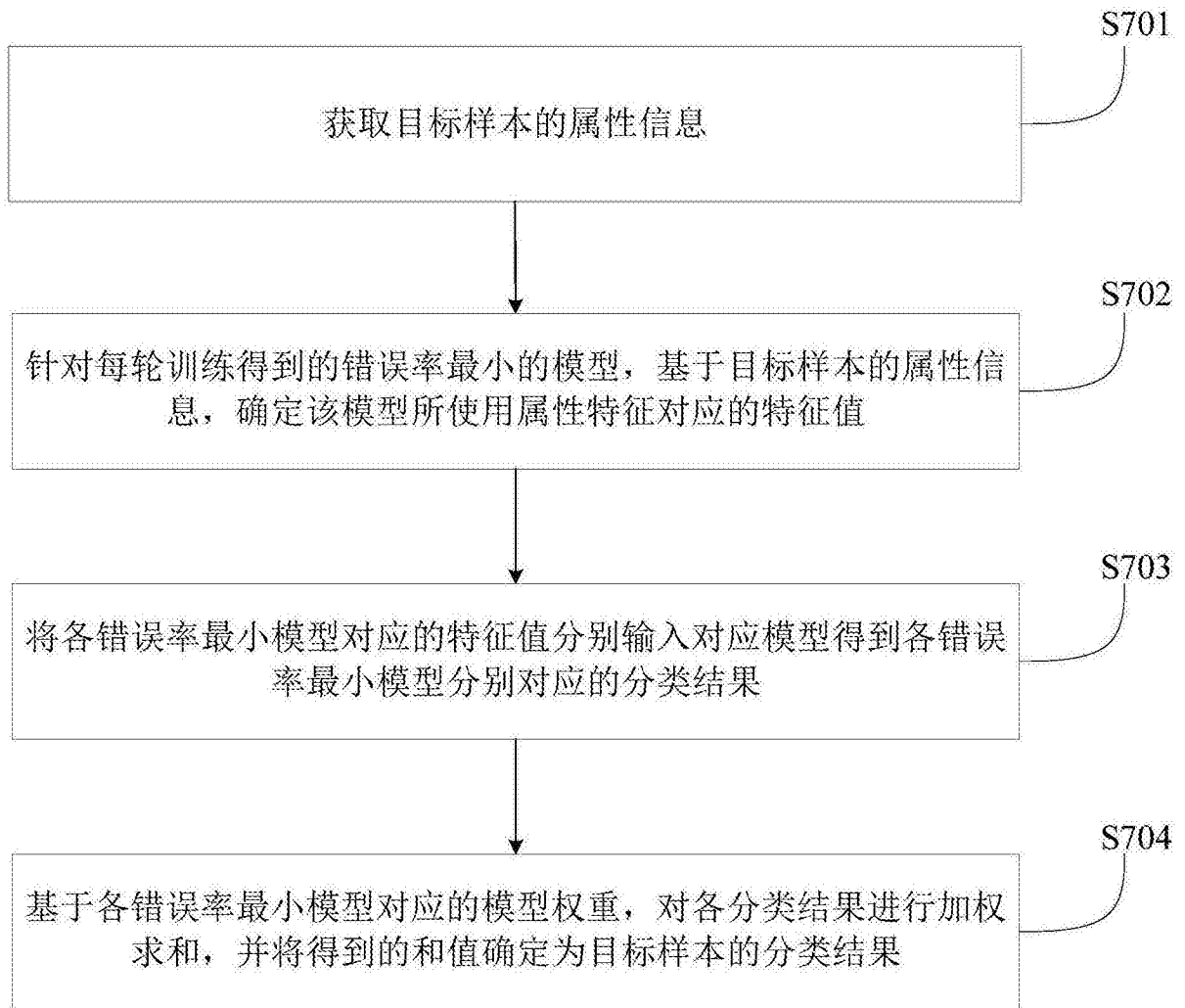


图7

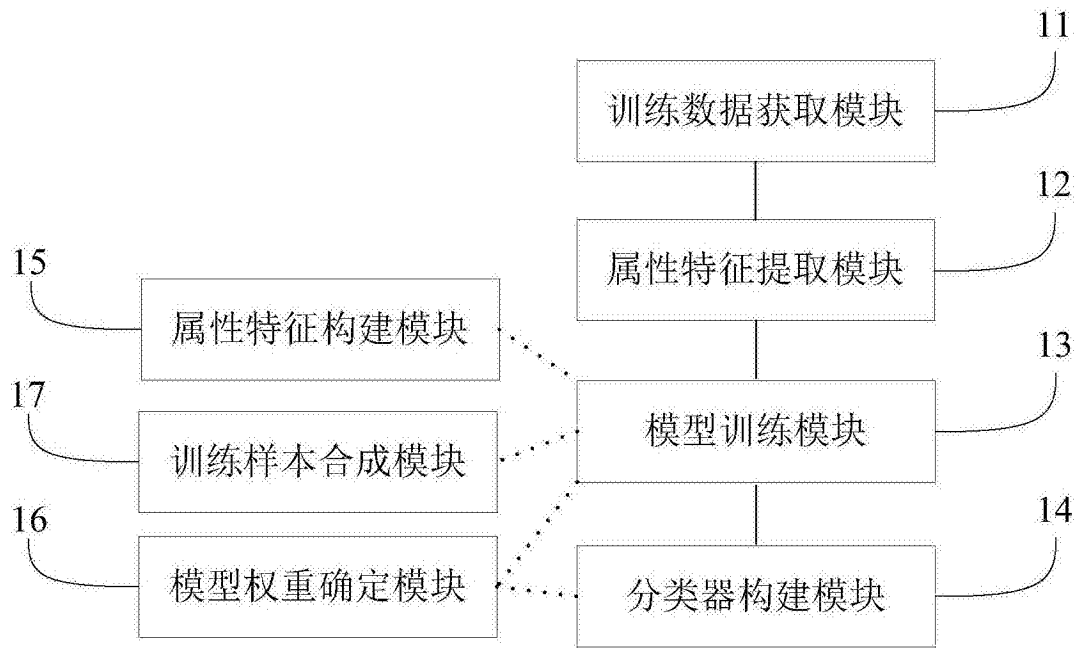


图8

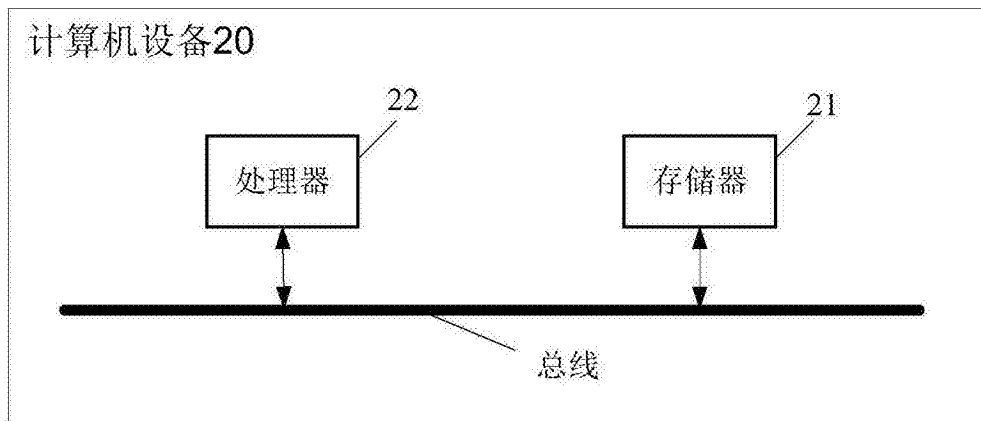


图9

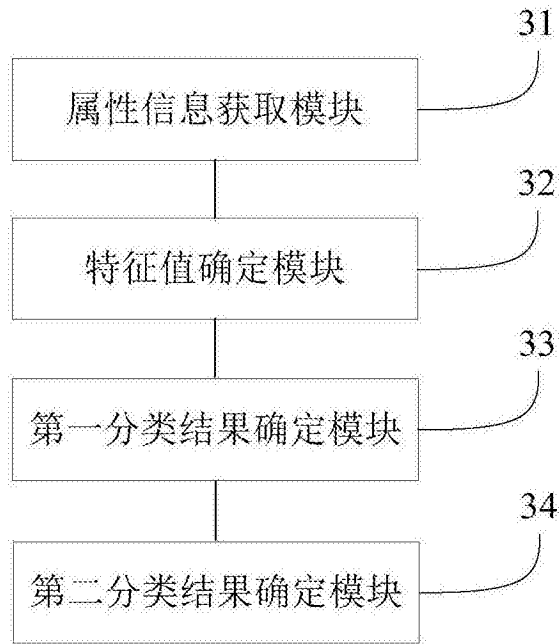


图10

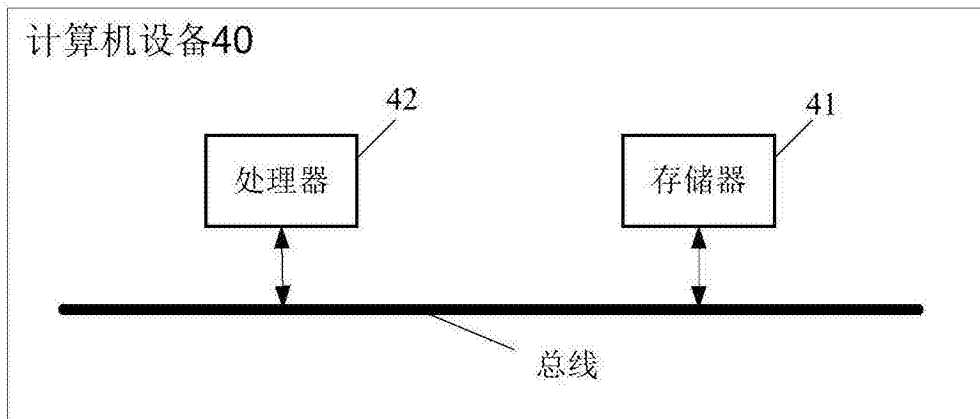


图11