(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 110175644 A
(43)申请公布日 2019.08.27

(21)申请号 201910445442.X

(22)申请日 2019.05.27

(71)申请人 恒安嘉新(北京)科技股份公司
地址 100098 北京市海淀区北三环西路25号27号楼五层5002室

(72)发明人 梁彧 郑祥波 金红 杨满智 刘长永 陈晓光

(74)专利代理机构 北京品源专利代理有限公司 11332
代理人 孟金喆

(51)Int.Cl.
G06K 9/62(2006.01)

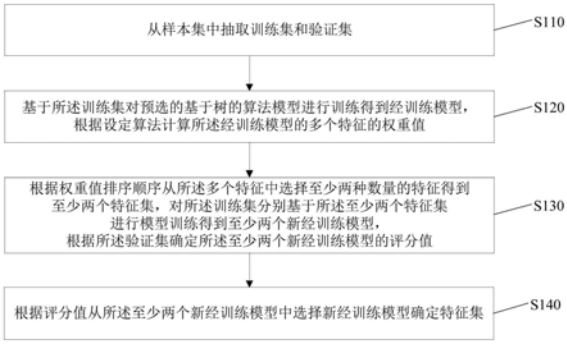
权利要求书2页 说明书10页 附图4页

(54)发明名称

特征选择方法、装置、电子设备、及存储介质

(57)摘要

本公开实施例公开了一种特征选择方法、装置、电子设备、及存储介质，方法包括：从样本集中抽取训练集和验证集；基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值；根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到至少两个特征集，对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型，根据所述验证集确定所述至少两个新经训练模型的评分值；根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。本分开实施例的技术方案能够自动根据所提供的样本集选出最佳特征集，能减少用户干预，提高用户的操作效率。



CN 110175644 A

1.一种特征选择方法,其特征在于,包括:

从样本集中抽取训练集和验证集;

基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型,根据设定算法计算所述经训练模型的多个特征的权重值;

根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到至少两个特征集,对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型,根据所述验证集确定所述至少两个新经训练模型的评分值;

根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。

2.根据权利要求1所述的方法,其特征在于,根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到至少两个特征集,对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型,根据所述验证集确定所述至少两个新经训练模型的评分值包括:

对所述多个特征按照权重值由大到小进行排序得到特征队列;

对所述训练集基于顺次从所述特征队列选择的至少一个特征进行模型训练得到新经训练模型,根据所述验证集确定所述新经训练模型的评分值;

依次增加选择的特征的数量重复执行上一步操作,得到至少两个新经训练模型、以及所述至少两个新经训练模型的评分值。

3.根据权利要求2所述的方法,其特征在于,依次增加选择的特征的数量重复执行上一步操作还包括:

在增加选择的特征的数量之后,判断所选择的特征数量是否大于设定阈值,若是则停止重复执行上一步操作的步骤。

4.根据权利要求2或3所述的方法,其特征在于,在每次重复执行上一步操作之后还包括:判断当前得到的新经训练模型的评分值与前一新经训练模型的评分值相比是否下降,若判断为下降且连续下降达到设定次数,则停止重复执行上一步操作的步骤。

5.根据权利要求1所述的方法,其特征在于,根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集包括:从所述至少两个新经训练模型中选择评分值最高的新经训练模型,确定所选择的新经训练模型所对应的特征集。

6.根据权利要求1所述的方法,其特征在于,所述基于树的算法模型包括随机森林、Adboost、GBDT、Xgboost、以及LightGBM。

7.根据权利要求1所述的方法,其特征在于,从样本集中抽取训练集和验证集包括:对所述样本集根据设定比例进行随机抽样确定训练集和验证集。

8.根据权利要求1所述的方法,其特征在于,根据设定算法计算所述经训练模型的多个特征的权重值包括:

根据在所述算法模型的所有树中作为划分属性的次数分别确定所述多个特征的权重值;或

根据在所述算法模型的所有树中作为划分属性时的平均信息熵的降低量分别确定所述多个特征的权重值;或

根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平均覆盖度分别确定所述多个特征的权重值;或

根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平方损失分别确定所述多个特征的权重值。

9.根据权利要求1所述的方法，其特征在于，所述特征选择方法用于建立机器学习模型，所述机器学习模型用于预测沉船上的乘客是否存活；

所述经训练模型的多个特征至少包括乘客的姓名、年龄、性别、所有船舱等级、上船的码头、终点码头、以及职业其中之一。

10.一种特征选择装置，其特征在于，包括：

训练集和验证集获取单元，用于从样本集中抽取训练集和验证集；

权重值计算单元，用于基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值；

模型训练及评分单元，用于根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到至少两个特征集，对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型，根据所述验证集确定所述至少两个新经训练模型的评分值；

特征确定单元，用于根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。

11.一种电子设备，其特征在于，包括：

一个或多个处理器；

存储器，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1-9中任一项所述方法的指令。

12.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该计算机程序被处理器执行时实现如权利要求1-9任一项所述方法的步骤。

特征选择方法、装置、电子设备、及存储介质

技术领域

[0001] 本公开实施例涉及机器学习技术领域，具体涉及一种特征选择方法、装置、电子设备、及存储介质。

背景技术

[0002] 在大数据人工智能时代，如何有效的从数据中挖掘出有价值的信息作为决策支持，显得越来越重要，而机器学习是解决这类问题的有效手段。关于机器学习相关问题的研究，已成为这个时代科技应用的热点。

[0003] 目前在实际机器学习场景中，比较流行的特征选择处理方式是采用人工选取，通过不断选取特征集，训练模型得到评分来选择较好的相关特征集。在理论上，特征选择的方法采用特征子集搜索机制与特征子集评价机制方法相结合而得到；特征子集搜索常用有“前向”搜索、“后向”搜索和“双向”搜索。在现有实现方案中往往采用穷举或递归的方法，其时间复杂度往往较高，计算耗时长，在特征备选集较大的情况下，往往难以实现。

发明内容

[0004] 有鉴于此，本公开实施例提供一种特征选择方法、装置、电子设备、及存储介质，以自动根据所提供的样本集选出最佳特征集，能减少用户干预。

[0005] 本公开实施例的其他特性和优点将通过下面的详细描述变得显然，或部分地通过本公开实施例的实践而习得。

[0006] 第一方面，本公开实施例提供了一种特征选择方法，包括：

[0007] 从样本集中抽取训练集和验证集；

[0008] 基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值；

[0009] 根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到至少两个特征集，对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型，根据所述验证集确定所述至少两个新经训练模型的评分值；

[0010] 根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。

[0011] 于一实施例中，根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到至少两个特征集，对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型，根据所述验证集确定所述至少两个新经训练模型的评分值包括：

[0012] 对所述多个特征按照权重值由大到小进行排序得到特征队列；

[0013] 对所述训练集基于顺次从所述特征队列选择的至少一个特征进行模型训练得到新经训练模型，根据所述验证集确定所述新经训练模型的评分值；

[0014] 依次增加选择的特征的数量重复执行上一步操作，得到至少两个新经训练模型、以及所述至少两个新经训练模型的评分值。

[0015] 于一实施例中，依次增加选择的特征的数量重复执行上一步操作还包括：

[0016] 在增加选择的特征的数量之后，判断所选择的特征数量是否大于设定阈值，若是则停止重复执行上一步操作的步骤。

[0017] 于一实施例中，在每次重复执行上一步操作之后还包括：判断当前得到的新经训练模型的评分值与前一新经训练模型的评分值相比是否下降，若判断为下降且连续下降达到设定次数，则停止重复执行上一步操作的步骤。

[0018] 于一实施例中，根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集包括：从所述至少两个新经训练模型中选择评分值最高的新经训练模型，确定所选择的新经训练模型所对应的特征集。

[0019] 于一实施例中，所述基于树的算法模型包括随机森林、Adboost、GBDT、Xgboost、以及LightGBM。

[0020] 于一实施例中，从样本集中抽取训练集和验证集包括：对所述样本集根据设定比例进行随机抽样确定训练集和验证集。

[0021] 于一实施例中，根据设定算法计算所述经训练模型的多个特征的权重值包括：

[0022] 根据在所述算法模型的所有树中作为划分属性的次数分别确定所述多个特征的权重值；或

[0023] 根据在所述算法模型的所有树中作为划分属性时的平均信息熵的降低量分别确定所述多个特征的权重值；或

[0024] 根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平均覆盖度分别确定所述多个特征的权重值；或

[0025] 根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平方损失分别确定所述多个特征的权重值。

[0026] 于一实施例中，所述特征选择方法用于建立机器学习模型，所述机器学习模型用于预测沉船上的乘客是否存活；所述经训练模型的多个特征至少包括乘客的姓名、年龄、性别、所有船舱等级、上船的码头、终点码头、以及职业其中之一。

[0027] 第二方面，本公开实施例还提供了一种特征选择装置，包括：

[0028] 训练集和验证集获取单元，用于从样本集中抽取训练集和验证集；

[0029] 权重值计算单元，用于基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值；

[0030] 模型训练及评分单元，用于根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到至少两个特征集，对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型，根据所述验证集确定所述至少两个新经训练模型的评分值；

[0031] 特征确定单元，用于根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。

[0032] 于一实施例中，所述模型训练及评分单元包括：

[0033] 排序子单元，用于对所述多个特征按照权重值由大到小进行排序得到特征队列；

[0034] 训练与评分子单元，用于对所述训练集基于顺次从所述特征队列选择的至少一个特征进行模型训练得到新经训练模型，根据所述验证集确定所述新经训练模型的评分值；

[0035] 循环选择子单元，用于依次增加选择的特征的数量重复执行所述训练与评分子单

元执行的操作，得到至少两个新经训练模型、以及所述至少两个新经训练模型的评分值。

[0036] 于一实施例中，所述循环选择子单元还用于：在增加选择的特征的数量之后，判断所选择的特征数量是否大于设定阈值，若是则停止重复执行上一步操作的步骤。

[0037] 于一实施例中，所述循环选择子单元还用于，在每次重复执行上一步操作之后，判断当前得到的新经训练模型的评分值与前一新经训练模型的评分值相比是否下降，若判断为下降且连续下降达到设定次数，则停止重复执行上一步操作的步骤。

[0038] 于一实施例中，所述特征确定单元用于：从所述至少两个新经训练模型中选择评分值最高的新经训练模型，确定所选择的新经训练模型所对应的特征集。

[0039] 于一实施例中，所述基于树的算法模型包括随机森林、Adboost、GBDT、Xgboost、以及LightGBM。

[0040] 于一实施例中，所述训练集和验证集获取单元用于：对所述样本集根据设定比例进行随机抽样确定训练集和验证集。

[0041] 于一实施例中，所述权重值计算单元用于：

[0042] 根据在所述算法模型的所有树中作为划分属性的次数分别确定所述多个特征的权重值；或

[0043] 根据在所述算法模型的所有树中作为划分属性时的平均信息熵的降低量分别确定所述多个特征的权重值；或

[0044] 根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平均覆盖度分别确定所述多个特征的权重值；或

[0045] 根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平方损失分别确定所述多个特征的权重值。

[0046] 于一实施例中，所述特征选择装置用于建立机器学习模型，所述机器学习模型用于预测沉船上的乘客是否存活；所述经训练模型的多个特征至少包括乘客的姓名、年龄、性别、所有船舱等级、上船的码头、终点码头、以及职业其中之一。

[0047] 第三方面，本公开实施例还提供了一种电子设备，包括：

[0048] 一个或多个处理器；

[0049] 存储器，用于存储一个或多个程序；

[0050] 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第一方面中任一项所述方法的指令。

[0051] 第四方面，本公开实施例还提供了一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如第一方面中任一项所述方法的步骤。

[0052] 本公开实施例的技术方案，通过从样本集中抽取训练集和验证集，基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算该经训练模型的多个特征的权重值，根据权重值排序顺序得到至少两个特征集分别进行模型训练和评分，根据评分确定特征集，能够在减少用户干预的情况下根据样本集选出最佳特征集，能够提高用户的操作效率。

附图说明

[0053] 为了更清楚地说明本公开实施例中的技术方案，下面将对本公开实施例描述中所

需要使用的附图作简单的介绍，显而易见地，下面描述中的附图仅仅是本公开实施例中的一部分实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据本公开实施例的内容和这些附图获得其他的附图。

[0054] 图1是本公开实施例提供的一种特征选择方法的流程示意图；

[0055] 图2是本公开实施例提供的另一种特征选择方法的流程示意图；

[0056] 图3是本公开实施例提供的一种特征选择装置的结构示意图；

[0057] 图4是本公开实施例提供的另一种特征选择装置的结构示意图；

[0058] 图5示出了适于用来实现本公开实施例的电子设备的结构示意图。

具体实施方式

[0059] 为使本公开实施例解决的技术问题、采用的技术方案和达到的技术效果更加清楚，下面将结合附图对本公开实施例的技术方案作进一步的详细描述，显然，所描述的实施例仅仅是本公开实施例中的一部分实施例，而不是全部的实施例。基于本公开实施例中的实施例，本领域技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本公开实施例保护的范围。

[0060] 需要说明的是，本公开实施例中术语“系统”和“网络”在本文中常被可互换使用。本公开实施例中提到的“和/或”是指”包括一个或更多个相关所列项目的任何和所有组合。本公开的说明书和权利要求书及附图中的术语“第一”、“第二”等是用于区别不同对象，而不是用于限定特定顺序。

[0061] 还需要说明是，本公开实施例中下述各个实施例可以单独执行，各个实施例之间也可以相互结合执行，本公开实施例对此不作具体限制。

[0062] 下面结合附图并通过具体实施方式来进一步说明本公开实施例的技术方案。

[0063] 图1示出了本公开实施例提供的一种特征选择方法的流程示意图，本实施例可适用于机器学习领域中监督学习的分类问题或回归问题，该方法可以由配置于计算机中的特征选择装置来执行，如图1所示，本实施例所述的特征选择方法包括：

[0064] 在步骤S110中，从样本集中抽取训练集和验证集。

[0065] 例如对所述样本集根据设定比例进行随机抽样确定训练集和验证集，具体地，例如计算原始样本集中样本数的70％值向上取整得到数目T；设定随机种子值，对原始样本集做洗牌操作后，取前T个样本作为训练集，剩余的作为验证集。

[0066] 在步骤S120中，基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值。

[0067] 所述基于树的算法模型包括多种，例如随机森林、Adboost、GBDT、Xgboost、以及LightGBM等，具体采用哪种基于树的算法模型可根据具体的应用领域、样本集的特征的特点等因素进行选择，本实施例对此并不作限定。

[0068] 需要说明的是，根据设定算法计算所述经训练模型的多个特征的权重值可采用多种方式计算，例如使用特征在所有树中作为划分属性的次数、在作为划分属性时平均信息熵的降低量、在作为划分属性时对样本的平均覆盖度、和在作为划分属性时对样本的平方损失等。

[0069] 具体地，例如可根据在所述算法模型的所有树中作为划分属性的次数分别确定所

述多个特征的权重值。

[0070] 又如，可根据在所述算法模型的所有树中作为划分属性时的平均信息熵的降低量分别确定所述多个特征的权重值。

[0071] 又如，还可根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平均覆盖度分别确定所述多个特征的权重值。

[0072] 再如，还可根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平方损失分别确定所述多个特征的权重值。

[0073] 在步骤S130中，根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到至少两个特征集，对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型，根据所述验证集确定所述至少两个新经训练模型的评分值。

[0074] 例如，对所述多个特征按照权重值由大到小进行排序得到特征队列；对所述训练集基于顺次从所述特征队列选择的至少一个特征进行模型训练得到新经训练模型，根据所述验证集确定所述新经训练模型的评分值；依次增加选择的特征的数量重复执行上一步操作，得到至少两个新经训练模型、以及所述至少两个新经训练模型的评分值。

[0075] 又如，依次增加选择的特征的数量重复执行上一步操作还可进一步包括：在增加选择的特征的数量之后，判断所选择的特征数量是否大于设定阈值，若是则停止重复执行上一步操作的步骤。

[0076] 进一步地，在每次重复执行上一步操作之后还可包括：判断当前得到的新经训练模型的评分值与前一新经训练模型的评分值相比是否下降，若判断为下降且连续下降达到设定次数，则停止重复执行上一步操作的步骤。

[0077] 上述具体步骤可使本实施例的具体方案在无需设定所选择的特征个数或比例的情况下，通过机器学习算法实现自动特征选择方法。另外，其对训练集选择出来的特征列表，在测试集中可以直接映射使用。

[0078] 在步骤S140中，根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。

[0079] 例如，可以从所述至少两个新经训练模型中选择评分值最高的新经训练模型，确定所选择的新经训练模型所对应的特征集。

[0080] 所述特征选择方法可用于建立机器学习模型，所述机器学习模型可用于预测沉船上的乘客是否存活；所述经训练模型的多个特征至少包括乘客的姓名、年龄、性别、所有船舱等级、上船的码头、终点码头、以及职业其中之一。

[0081] 本公开实施例的技术方案，通过从样本集中抽取训练集和验证集，基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算该经训练模型的多个特征的权重值，根据权重值排序顺序得到至少两个特征集分别进行模型训练和评分，根据评分确定特征集，能够在减少用户干预的情况下根据样本集选出最佳特征集，能够提高用户的操作效率。

[0082] 图2示出了本公开实施例提供的另一种特征选择方法的流程示意图，本实施例以前述实施例为基础，进行了改进优化。如图2所示，本实施例所述的特征选择方法包括：

[0083] 在步骤S210中，从样本集中抽取训练集和验证集。

[0084] 本步骤与上一实施例的步骤S110相同，本实施例在此不作赘述。

[0085] 在步骤S220中，基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值。

[0086] 本步骤与上一实施例的步骤S120相同，本实施例在此不作赘述。

[0087] 在步骤S230中，对所述多个特征按照权重值由大到小进行排序得到特征队列。

[0088] 例如可对特征依据重要性进行由大到小排序，创建一个新的空列表接收特征集。

[0089] 在步骤S240中，对所述训练集基于顺次从所述特征队列选择的至少一个特征进行模型训练得到新经训练模型，根据所述验证集确定所述新经训练模型的评分值。

[0090] 在步骤S250中，依次增加选择的特征的数量重复执行上一步操作，得到至少两个新经训练模型、以及所述至少两个新经训练模型的评分值。

[0091] 例如在每一次增加选择的特征的数量之后，判断所选择的特征数量是否大于设定阈值，若是则停止重复执行上一步操作S240的步骤。

[0092] 又例如，还可在每次重复执行上一步操作之后还判断当前得到的新经训练模型的评分值与前一新经训练模型的评分值相比是否下降，若判断为下降且连续下降达到设定次数，则停止重复执行上一步操作的步骤。

[0093] 具体地，例如从排序后的特征列表中依特征重要性大小逐步选择一个特征添加到新列表中，生成序列特征子集；每添加一次训练一个新的模型并在验证集上得到评分，这样就得到与序列特征子集对应的评分列表。

[0094] 更具体地，可由一个空列表一次添加一个特征生成特征集，训练一个模型对验证集测试，得到评分，其评分指标通常是分类或回归问题的评价指标中的一种。为了加快选择速度，可以加入压缩初始特征空间和早停技巧。例如可设定前S(S为大于或等于1的整数)个特征重要性值对应的特征子集作为初始特征集，压缩了初始特征空间。进一步地，还可设定早停值k(k为大于或等于1的整数)，即当连续添加k次特征，训练出的模型对验证集测试评分无变化，则停止继续添加特征，停止训练模型，生成的评分列表作为最终的评分列表。

[0095] 在步骤S260中，根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。

[0096] 例如，可以从所述至少两个新经训练模型中选择评分值最高的新经训练模型，确定所选择的新经训练模型所对应的特征集。

[0097] 例如，所述特征选择方法可用于建立机器学习模型。例如所述机器学习模型可用于预测沉船上的乘客是否存活，所述经训练模型的多个特征至少包括乘客的姓名、年龄、性别、所有船舱等级、上船的码头、终点码头、以及职业其中之一。

[0098] 本公开实施例的技术方案在上一实施例的基础之上进一步具体地示例了自动选择特征集的方法，可以在不主观设定所要选择的特征个数和比例的情况下，自动根据所提供的数据选出最佳特征子集，减少人为干预，客观性强且结果更具有稳定性，实现简单，效率高。

[0099] 更进一步地，本实施例还进一步增加了设置早停的方法示例，进一步地增强了计算的灵活性，大大宿减了时间复杂度，提高了运算效率。

[0100] 图3示出了本公开实施例提供的一种特征选择装置的结构示意图，如图3所示，本实施例所述的特征选择装置包括训练集和验证集获取单元310、权重值计算单元320、模型训练及评分单元330、以及特征确定单元340。

[0101] 所述训练集和验证集获取单元310被配置为，用于从样本集中抽取训练集和验证集；

[0102] 所述权重值计算单元320被配置为，用于基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值；

[0103] 所述模型训练及评分单元330被配置为，用于根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到至少两个特征集，对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型，根据所述验证集确定所述至少两个新经训练模型的评分值；

[0104] 所述特征确定单元340被配置为，用于根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。

[0105] 进一步地，所述特征确定单元340被配置为，用于从所述至少两个新经训练模型中选择评分值最高的新经训练模型，确定所选择的新经训练模型所对应的特征集。

[0106] 进一步地，所述基于树的算法模型包括随机森林、Adboost、GBDT、Xgboost、以及LightGBM。

[0107] 进一步地，所述训练集和验证集获取单元310被配置为，用于对所述样本集根据设定比例进行随机抽样确定训练集和验证集。

[0108] 进一步地，所述权重值计算单元320被配置为，用于：

[0109] 根据在所述算法模型的所有树中作为划分属性的次数分别确定所述多个特征的权重值；或

[0110] 根据在所述算法模型的所有树中作为划分属性时的平均信息熵的降低量分别确定所述多个特征的权重值；或

[0111] 根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平均覆盖度分别确定所述多个特征的权重值；或

[0112] 根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平方损失分别确定所述多个特征的权重值。

[0113] 例如，所述特征选择装置可用于建立机器学习模型。例如所述机器学习模型可用于预测沉船上的乘客是否存活，所述经训练模型的多个特征至少包括乘客的姓名、年龄、性别、所有船舱等级、上船的码头、终点码头、以及职业其中之一。

[0114] 本实施例提供的特征选择装置可执行本公开实施例方法实施例所提供的特征选择方法，具备执行方法相应的功能模块和有益效果。

[0115] 图4示出了本公开实施例提供的另一种特征选择装置的结构示意图，如图4所示，本实施例所述的特征选择装置包括训练集和验证集获取单元410、权重值计算单元420、模型训练及评分单元430、以及特征确定单元440，其中，所述模型训练及评分单元430包括排序子单元431、训练与评分子单元432、以及循环选择子单元433。

[0116] 所述训练集和验证集获取单元410被配置为，用于从样本集中抽取训练集和验证集；

[0117] 所述权重值计算单元420被配置为，用于基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值；

[0118] 所述排序子单元431被配置为，用于对所述多个特征按照权重值由大到小进行排

序得到特征队列；

[0119] 所述训练与评分子单元432被配置为，用于对所述训练集基于顺次从所述特征队列选择的至少一个特征进行模型训练得到新经训练模型，根据所述验证集确定所述新经训练模型的评分值；

[0120] 所述循环选择子单元433被配置为，用于依次增加选择的特征的数量重复执行所述训练与评分子单元执行的操作，得到至少两个新经训练模型、以及所述至少两个新经训练模型的评分值。

[0121] 所述特征确定单元440被配置为，用于根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。

[0122] 进一步地，所述循环选择子单元433被配置为，还用于：在增加选择的特征的数量之后，判断所选择的特征数量是否大于设定阈值，若是则停止重复执行所述训练与评分子单元432操作的步骤。

[0123] 进一步地，所述循环选择子单元433被配置为，还用于在每次重复执行上一步操作之后，判断当前得到的新经训练模型的评分值与前一新经训练模型的评分值相比是否下降，若判断为下降且连续下降达到设定次数，则停止重复执行上一步操作的步骤。

[0124] 进一步地，所述特征确定单元440被配置为，用于从所述至少两个新经训练模型中选择评分值最高的新经训练模型，确定所选择的新经训练模型所对应的特征集。

[0125] 进一步地，所述基于树的算法模型包括随机森林、Adboost、GBDT、Xgboost、以及LightGBM。

[0126] 进一步地，所述训练集和验证集获取单元410被配置为，用于对所述样本集根据设定比例进行随机抽样确定训练集和验证集。

[0127] 进一步地，所述权重值计算单元420被配置为，用于：

[0128] 根据在所述算法模型的所有树中作为划分属性的次数分别确定所述多个特征的权重值；或

[0129] 根据在所述算法模型的所有树中作为划分属性时的平均信息熵的降低量分别确定所述多个特征的权重值；或

[0130] 根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平均覆盖度分别确定所述多个特征的权重值；或

[0131] 根据在所述算法模型的所有树中作为划分属性时对所述训练样本的平方损失分别确定所述多个特征的权重值。

[0132] 例如，所述特征选择装置可用于建立机器学习模型。例如所述机器学习模型可用于预测沉船上的乘客是否存活，所述经训练模型的多个特征至少包括乘客的姓名、年龄、性别、所有船舱等级、上船的码头、终点码头、以及职业其中之一。

[0133] 本实施例提供的特征选择装置可执行本公开实施例方法实施例所提供的特征选择方法，具备执行方法相应的功能模块和有益效果。

[0134] 下面参考图5，其示出了适于用来实现本公开实施例的电子设备500的结构示意图。本公开实施例中的终端设备可以包括但不限于诸如移动电话、笔记本电脑、数字广播接收器、PDA(个人数字助理)、PAD(平板电脑)、PMP(便携式多媒体播放器)、车载终端(例如车载导航终端)等等的移动终端以及诸如数字TV、台式计算机等等的固定终端。图5示出的电

子设备仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

[0135] 如图5所示，电子设备500可以包括处理装置(例如中央处理器、图形处理器等)501，其可以根据存储在只读存储器(ROM)502中的程序或者从存储装置508加载到随机访问存储器(RAM)503中的程序而执行各种适当的动作和处理。在RAM 503中，还存储有电子设备500操作所需的各种程序和数据。处理装置501、ROM 502以及RAM 503通过总线504彼此相连。输入/输出(I/O)接口505也连接至总线504。

[0136] 通常，以下装置可以连接至I/O接口505：包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置506；包括例如液晶显示器(LCD)、扬声器、振动器等的输出装置507；包括例如磁带、硬盘等的存储装置508；以及通信装置509。通信装置509可以允许电子设备500与其他设备进行无线或有线通信以交换数据。虽然图5示出了具有各种装置的电子设备500，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。

[0137] 特别地，根据本公开实施例的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开实施例的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信装置509从网络上被下载和安装，或者从存储装置508被安装，或者从ROM 502被安装。在该计算机程序被处理装置501执行时，执行本公开实施例的方法中限定的上述功能。

[0138] 需要说明的是，本公开实施例上述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开实施例中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开实施例中，计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：电线、光缆、RF(射频)等等，或者上述的任意合适的组合。

[0139] 上述计算机可读介质可以是上述电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。

[0140] 上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该电子设备执行时，使得该电子设备：从样本集中抽取训练集和验证集；

[0141] 基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值；根据权重值排序顺序从所述多个特征中选

择至少两种数量的特征得到至少两个特征集，对所述训练集分别基于所述至少两个特征集进行模型训练得到至少两个新经训练模型，根据所述验证集确定所述至少两个新经训练模型的评分值；根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集。

[0142] 可以以一种或多种程序设计语言或其组合来编写用于执行本公开实施例的操作的计算机程序代码，上述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。

[0143] 附图中的流程图和框图，图示了按照本公开实施例各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0144] 描述于本公开实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。其中，单元的名称在某种情况下并不构成对该单元本身的限定，例如，第一获取单元还可以被描述为“获取至少两个网际协议地址的单元”。

[0145] 以上描述仅为本公开实施例的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本公开实施例中所涉及的公开范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述公开构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开实施例中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

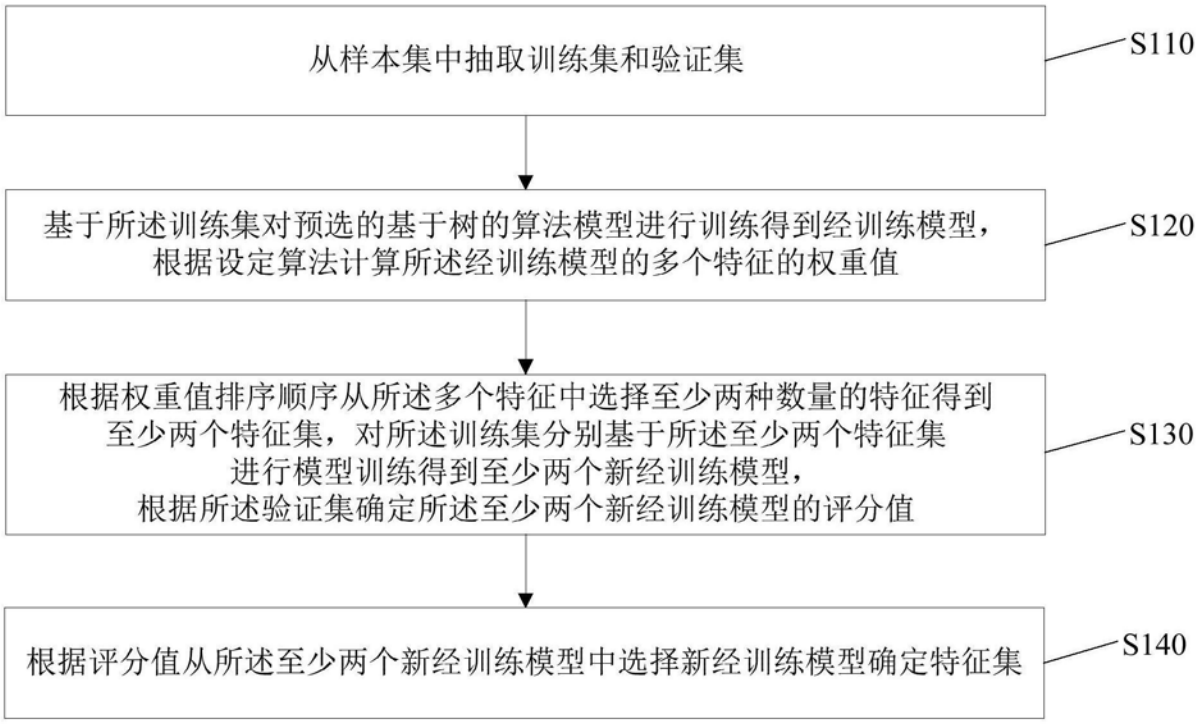

从样本集中抽取训练集和验证集
S110
基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，
根据设定算法计算所述经训练模型的多个特征的权重值
S120
根据权重值排序顺序从所述多个特征中选择至少两种数量的特征得到
至少两个特征集，对所述训练集分别基于所述至少两个特征集
进行模型训练得到至少两个新经训练模型，
根据所述验证集确定所述至少两个新经训练模型的评分值
S130
根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集
S140

图1

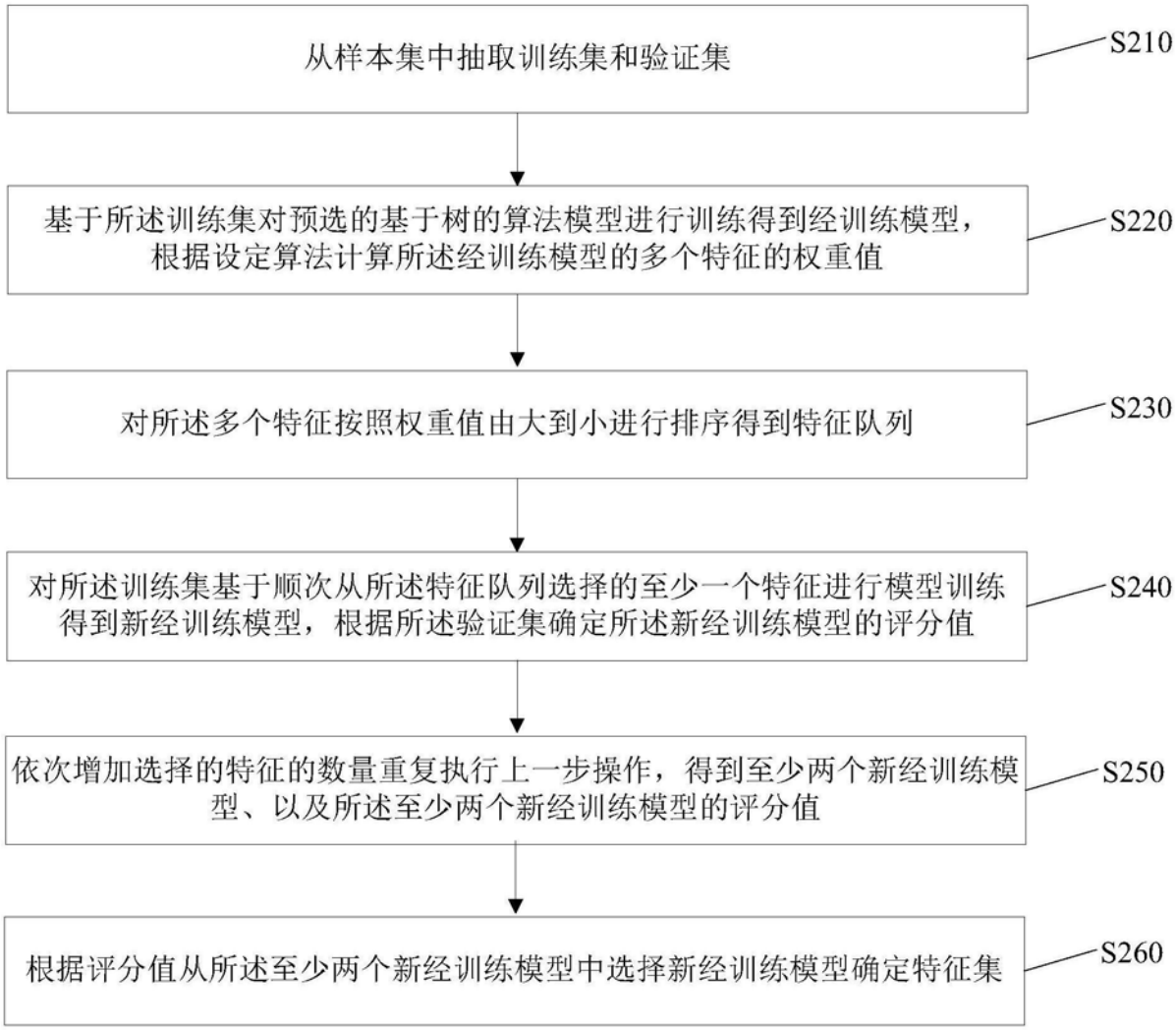

从样本集中抽取训练集和验证集
S210
基于所述训练集对预选的基于树的算法模型进行训练得到经训练模型，根据设定算法计算所述经训练模型的多个特征的权重值
S220
对所述多个特征按照权重值由大到小进行排序得到特征队列
S230
对所述训练集基于顺次从所述特征队列选择的至少一个特征进行模型训练得到新经训练模型，根据所述验证集确定所述新经训练模型的评分值
S240
依次增加选择的特征的数量重复执行上一步操作，得到至少两个新经训练模型、以及所述至少两个新经训练模型的评分值
S250
根据评分值从所述至少两个新经训练模型中选择新经训练模型确定特征集
S260

图2

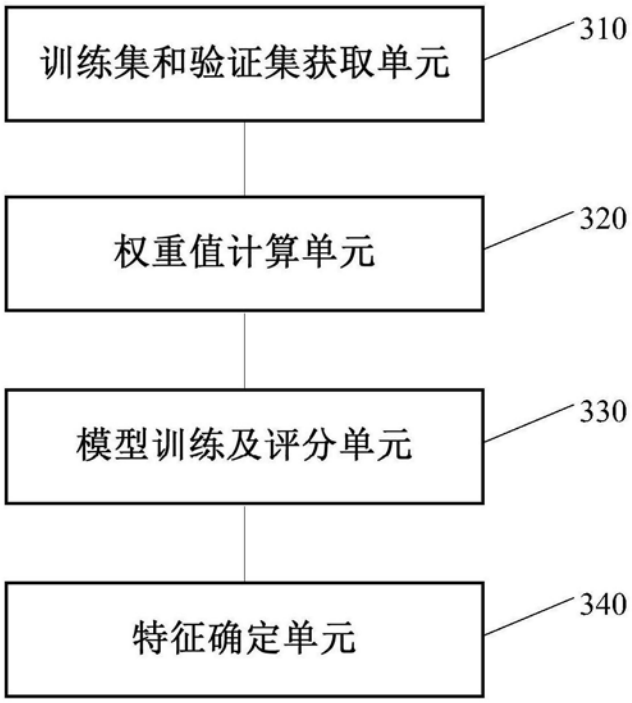


图3

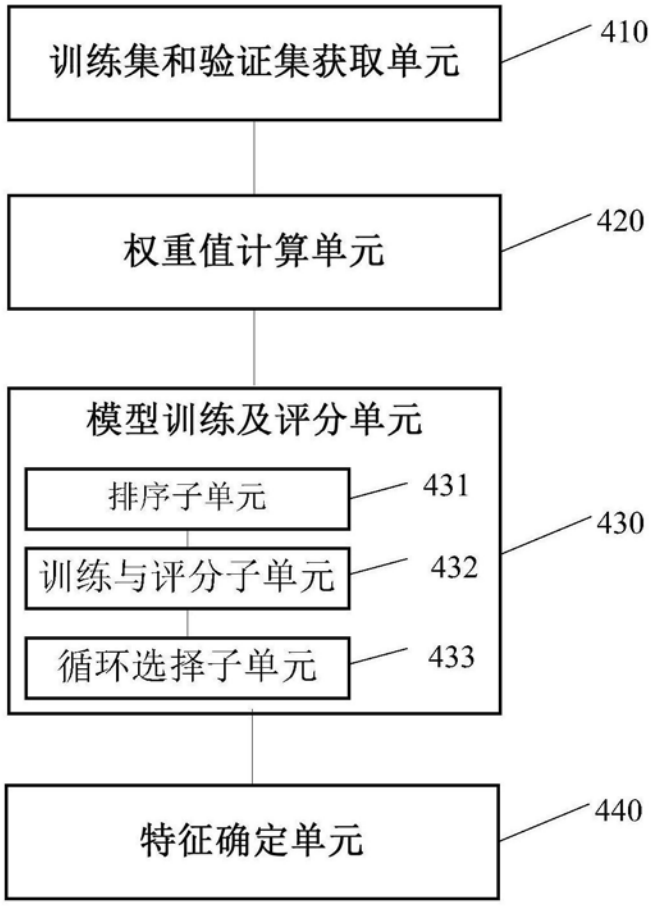

训练集和验证集获取单元
410
权重值计算单元
420
模型训练及评分单元
排序子单元
431
训练与评分子单元
432
循环选择子单元
433
430
特征确定单元
440

图4

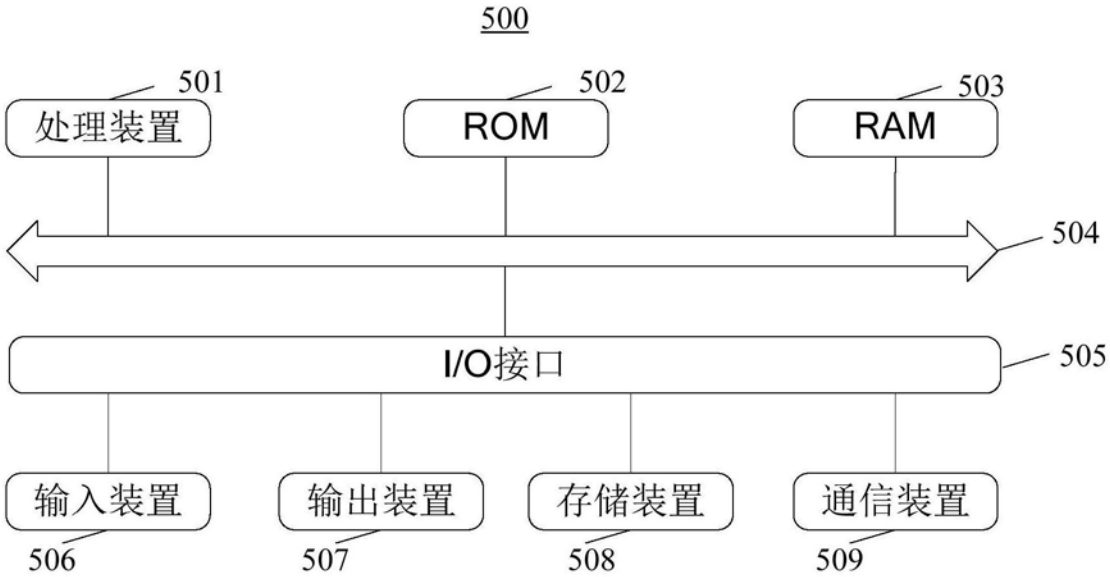

500
501
处理装置
502
ROM
503
RAM
504
I/O接口
505
输入装置
506
输出装置
507
存储装置
508
通信装置
509

图5