(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 108021984 A
(43)申请公布日 2018.05.11

(21)申请号 201610935697.0

(22)申请日 2016.11.01

(71)申请人 第四范式(北京)技术有限公司
地址 100085 北京市海淀区上地东路35号颐泉汇大厦写字楼A座610室

(72)发明人 罗远飞 涂威威

(74)专利代理机构 北京展翼知识产权代理事务所(特殊普通合伙) 11452
代理人 屠长存

(51)Int.Cl.
G06N 99/00(2010.01)
G06K 9/62(2006.01)

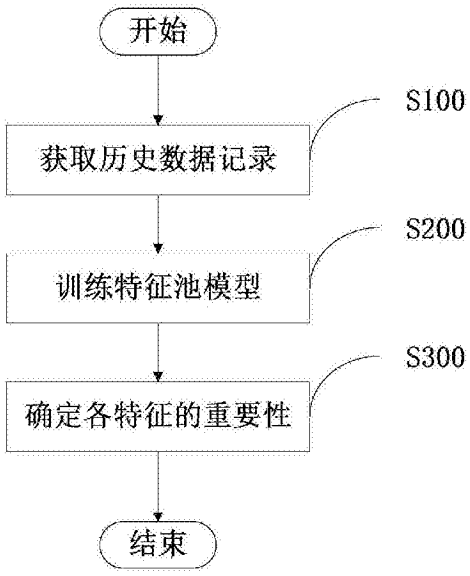
权利要求书2页 说明书23页 附图3页

(54)发明名称

确定机器学习样本的特征重要性的方法及系统

(57)摘要

提供了一种确定机器学习样本的特征重要性的方法及系统，所述方法包括：(A)获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和至少一个属性信息；(B)利用获取的历史数据记录，训练至少一个特征池模型，其中，特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型；(C)获取所述至少一个特征池模型的效果，并根据获取的所述至少一个特征池模型的效果来确定所述各个特征的重要性，其中，在步骤(B)中，通过对所述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。通过所述方法和系统，可有效确定机器学习样本中各个特征的重要性。



CN 108021984 A

1.一种确定机器学习样本的各个特征的重要性的方法,包括:

(A)获取历史数据记录,其中,所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息;

(B)利用获取的历史数据记录,训练至少一个特征池模型,其中,特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型;

(C)获取所述至少一个特征池模型的效果,并根据获取的所述至少一个特征池模型的效果来确定所述各个特征的重要性,

其中,在步骤(B)中,通过对所述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。

2.如权利要求1所述的方法,其中,在步骤(C)中,根据特征池模型在原始测试数据集和变换测试数据集上的效果之间的差异来确定所述特征池模型所基于的相应特征的重要性,

其中,变换测试数据集是指通过对原始测试数据集中的其重要性待确定的目标特征的取值替换为以下项之一而获得的数据集:零值、随机数值、通过将目标特征的原始取值扰乱顺序后得到的值。

3.如权利要求1所述的方法,其中,所述至少一个特征池模型包括多个基于不同特征组来提供关于机器学习问题的预测结果的机器学习模型,

其中,在步骤(C)中,根据所述至少一个特征池模型在原始测试数据集上的效果之间的差异来确定所述各个特征的重要性。

4.如权利要求3所述的方法,其中,所述至少一个特征池模型包括一个或多个主特征池模型以及分别与每个主特征池模型相应的至少一个子特征池模型,其中,子特征池模型是指基于与其相应的主特征池模型所基于的特征之中除了其重要性待确定的目标特征之外的剩余特征来提供关于机器学习问题的预测结果的机器学习模型,

其中,在步骤(C)中,根据主特征池模型和与其相应的各个子特征池模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

5.如权利要求3所述的方法,其中,所述至少一个特征池模型包括多个单特征模型,其中,单特征模型是指基于所述各个特征之中的其重要性待确定的目标特征来提供关于机器学习问题的预测结果的机器学习模型,

其中,在步骤(C)中,根据单特征模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

6.如权利要求1所述的方法,其中,所述离散化运算包括基本分箱运算和至少一个附加运算。

7.如权利要求6所述的方法,其中,所述至少一个附加运算包括与基本分箱运算分箱方式相同但分箱参数不同的附加分箱运算;或者,所述至少一个附加运算包括与基本分箱运算分箱方式不同的附加分箱运算。

8.如权利要求1所述的方法,其中,步骤(B)还包括:向用户提供用于配置特征池模型的以下项目之中的至少一个项目的界面:特征池模型所基于的至少一部分特征、特征池模型的算法种类、特征池模型的算法参数、离散化运算的运算种类、离散化运算的运算参数,

并且,在步骤(B)中,根据用户通过所述界面配置的项目来分别训练特征池模型。

9.一种确定机器学习样本的各个特征的重要性的系统，包括：

数据记录获取装置，用于获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息；

模型训练装置，用于利用获取的历史数据记录，训练至少一个特征池模型，其中，特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型；

重要性确定装置，用于获取所述至少一个特征池模型的效果，并根据获取的所述至少一个特征池模型的效果来确定所述各个特征的重要性，

其中，模型训练装置通过对所述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。

10.一种确定机器学习样本的各个特征的重要性的计算装置，包括存储部件和处理器，存储部件中存储有计算机可执行指令集合，当所述计算机可执行指令集合被所述处理器执行时，执行下述步骤：

(A)获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息；

(B)利用获取的历史数据记录，训练至少一个特征池模型，其中，特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型；

(C)获取所述至少一个特征池模型的效果，并根据获取的所述至少一个特征池模型的效果来确定所述各个特征的重要性，

其中，在步骤(B)中，通过对所述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。

确定机器学习样本的特征重要性的方法及系统

技术领域

[0001] 本发明总体说来涉及人工智能领域，更具体地说，涉及一种针对机器学习样本的特征重要性确定方法及系统。

背景技术

[0002] 随着海量数据的出现，人工智能技术得到了迅速发展，而为了从大量数据中挖掘出价值，需要基于数据记录来产生适用于机器学习的样本。

[0003] 这里，每条数据记录可被看做关于一个事件或对象的描述，对应于一个示例或样例。在数据记录中，包括反映事件或对象在某方面的表现或性质的各个事项，这些事项可称为“属性”。

[0004] 实践中，机器学习模型的预测效果与模型的选择、可用的数据和特征的提取等有关。如何从原始数据记录的各个属性提取出机器学习样本的特征，将会对机器学习模型的效果带来很大的影响。相应地，不论从模型训练还是模型理解的角度来看，都很需要获知机器学习样本的各个特征的重要程度。例如，可根据基于XGBoost训练出的树模型，计算每个特征的期望分裂增益，然后计算特征重要性。上述方式虽然能考虑特征之间的相互作用，但训练代价高，且不同参数对特征重要性的影响较大。

[0005] 实际上，特征的重要性难以直观确定，往往需要技术人员不仅掌握机器学习的知识，还需要对实际预测问题有深入的理解，而预测问题往往结合着不同行业的不同实践经验，导致很难达到满意的效果。

发明内容

[0006] 本发明的示例性实施例旨在克服现有技术中难以有效地确定机器学习样本的各个特征的重要性的缺陷。

[0007] 根据本发明的示例性实施例，提供一种确定机器学习样本的各个特征的重要性的方法，包括：(A)获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息；(B)利用获取的历史数据记录，训练至少一个特征池模型，其中，特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型；(C)获取所述至少一个特征池模型的效果，并根据获取的所述至少一个特征池模型的效果来确定所述各个特征的重要性，其中，在步骤(B)中，通过对所述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。

[0008] 可选地，在所述方法中，在步骤(C)中，根据特征池模型在原始测试数据集和变换测试数据集上的效果之间的差异来确定所述特征池模型所基于的相应特征的重要性，其中，变换测试数据集是指通过对原始测试数据集中的其重要性待确定的目标特征的取值替换为以下项之一而获得的数据集：零值、随机数值、通过将目标特征的原始取值扰乱顺序后得到的值。

[0009] 可选地，在所述方法中，所述至少一个特征池模型包括一个全部特征模型，其中，全部特征模型是指基于所述各个特征之中的全部特征来提供关于机器学习问题的预测结果的机器学习模型。

[0010] 可选地，在所述方法中，所述至少一个特征池模型包括多个基于不同特征组来提供关于机器学习问题的预测结果的机器学习模型，其中，在步骤(C)中，根据所述至少一个特征池模型在原始测试数据集上的效果之间的差异来确定所述各个特征的重要性。

[0011] 可选地，在所述方法中，所述至少一个特征池模型包括一个或多个主特征池模型以及分别与每个主特征池模型相应的至少一个子特征池模型，其中，子特征池模型是指基于与其相应的主特征池模型所基于的特征之中除了其重要性待确定的目标特征之外的剩余特征来提供关于机器学习问题的预测结果的机器学习模型，其中，在步骤(C)中，根据主特征池模型和与其相应的各个子特征池模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

[0012] 可选地，在所述方法中，所述至少一个特征池模型包括多个单特征模型，其中，单特征模型是指基于所述各个特征之中的其重要性待确定的目标特征来提供关于机器学习问题的预测结果的机器学习模型，其中，在步骤(C)中，根据单特征模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

[0013] 可选地，在所述方法中，所述离散化运算包括基本分箱运算和至少一个附加运算。

[0014] 可选地，在所述方法中，所述至少一个附加运算包括以下种类的运算之中的至少一种运算：对数运算、指数运算、绝对值运算、高斯变换运算。

[0015] 可选地，在所述方法中，所述至少一个附加运算包括与基本分箱运算分箱方式相同但分箱参数不同的附加分箱运算；或者，所述至少一个附加运算包括与基本分箱运算分箱方式不同的附加分箱运算。

[0016] 可选地，在所述方法中，基本分箱运算和附加分箱运算分别对应于不同宽度的等宽分箱运算或不同深度的等深分箱。

[0017] 可选地，在所述方法中，所述不同宽度或不同深度在数值上构成等比数列或等差数列。

[0018] 可选地，在所述方法中，执行基本分箱运算和/或附加分箱运算的步骤包括：额外设置离群箱，使得具有离群值的连续特征被分到所述离群箱。

[0019] 可选地，在所述方法中，在步骤(B)中，基于对数几率回归(logistic regressive)算法来训练特征池模型。

[0020] 可选地，在所述方法中，特征池模型的效果包括特征池模型的AUC。

[0021] 可选地，在所述方法中，所述原始测试数据集由获取的历史数据记录构成，其中，在步骤(B)中，将获取的历史数据记录划分为多组历史数据记录以逐步地训练各个特征池模型，并且，步骤(B)还包括：使用经过当前组历史数据记录训练后的特征池模型来针对下一组历史数据记录执行预测以得到与所述下一组历史数据记录相应的分组AUC，并综合各个分组AUC来得到特征池模型的AUC，其中，在得到与所述下一组历史数据记录相应的分组AUC之后，利用所述下一组历史数据记录来继续训练经过所述当前组历史数据记录训练后的特征池模型。

[0022] 可选地，在所述方法中，在步骤(B)中，在使用经过当前组历史数据记录训练后的

特征池模型来针对下一组历史数据记录执行预测时，当所述下一组历史数据记录包括缺少用于产生特征池模型所基于的至少一部分特征的属性信息的缺失历史数据记录时，基于以下处理之一来得到与所述下一组历史数据记录相应的分组AUC：仅利用所述下一组历史数据记录中除了缺失历史数据记录以外的其他历史数据记录的预测结果来计算分组AUC；利用所述下一组历史数据记录的全部历史数据记录的预测结果来计算分组AUC，其中，将缺失历史数据记录的预测结果设置为默认值，所述默认值基于预测结果的取值范围来确定或基于获取的历史数据记录的标记分布来确定；将利用所述下一组历史数据记录中除了缺失历史数据记录以外的其他历史数据记录的预测结果计算出的AUC与所述其他历史数据记录在所述下一组历史数据记录中所占的比例相乘来得到分组AUC。

[0023] 可选地，在所述方法中，在步骤(B)中，在基于对数几率回归算法来训练特征池模型时，针对连续特征设置的正则项不同于针对非连续特征设置的正则项。

[0024] 可选地，在所述方法中，步骤(B)还包括：向用户提供用于配置特征池模型的以下项目之中的至少一个项目的界面：特征池模型所基于的至少一部分特征、特征池模型的算法种类、特征池模型的算法参数、离散化运算的运算种类、离散化运算的运算参数，并且，在步骤(B)中，根据用户通过所述界面配置的项目来分别训练特征池模型。

[0025] 可选地，在所述方法中，在步骤(B)中，响应于用户关于确定特征重要性的指示来向用户提供所述界面。

[0026] 可选地，所述方法还包括：(D)以图形化方式向用户展示确定的各个特征的重要性。

[0027] 可选地，在所述方法中，在步骤(D)中，按照特征的重要性的顺序来展示各个特征，并且/或者，对所述各个特征之中的一部分特征进行突出显示，其中，所述一部分特征包括与高重要性对应的重要特征、与低重要性对应的不重要特征和/或与异常重要性对应的异常特征。

[0028] 根据本发明的另一示例性实施例，提供一种确定机器学习样本的各个特征的重要性的系统，包括：数据记录获取装置，用于获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息；模型训练装置，用于利用获取的历史数据记录，训练至少一个特征池模型，其中，特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型；重要性确定装置，用于获取所述至少一个特征池模型的效果，并根据获取的所述至少一个特征池模型的效果来确定所述各个特征的重要性，其中，模型训练装置通过对所述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。

[0029] 可选地，在所述系统中，重要性确定装置根据特征池模型在原始测试数据集和变换测试数据集上的效果之间的差异来确定所述特征池模型所基于的相应特征的重要性，其中，变换测试数据集是指通过对原始测试数据集中的其重要性待确定的目标特征的取值替换为以下项之一而获得的数据集：零值、随机数值、通过将目标特征的原始取值扰乱顺序后得到的值。

[0030] 可选地，在所述系统中，所述至少一个特征池模型包括一个全部特征模型，其中，全部特征模型是指基于所述各个特征之中的全部特征来提供关于机器学习问题的预测结果的机器学习模型。

[0031] 可选地，在所述系统中，所述至少一个特征池模型包括多个基于不同特征组来提供关于机器学习问题的预测结果的机器学习模型，其中，重要性确定装置根据所述至少一个特征池模型在原始测试数据集上的效果之间的差异来确定所述各个特征的重要性。

[0032] 可选地，在所述系统中，所述至少一个特征池模型包括一个或多个主特征池模型以及分别与每个主特征池模型相应的至少一个子特征池模型，其中，子特征池模型是指基于与其相应的主特征池模型所基于的特征之中除了其重要性待确定的目标特征之外的剩余特征来提供关于机器学习问题的预测结果的机器学习模型，其中，重要性确定装置根据主特征池模型和与其相应的各个子特征池模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

[0033] 可选地，在所述系统中，所述至少一个特征池模型包括多个单特征模型，其中，单特征模型是指基于所述各个特征之中的其重要性待确定的目标特征来提供关于机器学习问题的预测结果的机器学习模型，其中，重要性确定装置根据单特征模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

[0034] 可选地，在所述系统中，所述离散化运算包括基本分箱运算和至少一个附加运算。

[0035] 可选地，在所述系统中，所述至少一个附加运算包括以下种类的运算之中的至少一种运算：对数运算、指数运算、绝对值运算、高斯变换运算。

[0036] 可选地，在所述系统中，所述至少一个附加运算包括与基本分箱运算分箱方式相同但分箱参数不同的附加分箱运算；或者，所述至少一个附加运算包括与基本分箱运算分箱方式不同的附加分箱运算。

[0037] 可选地，在所述系统中，基本分箱运算和附加分箱运算分别对应于不同宽度的等宽分箱运算或不同深度的等深分箱。

[0038] 可选地，在所述系统中，所述不同宽度或不同深度在数值上构成等比数列或等差数列。

[0039] 可选地，在所述系统中，执行基本分箱运算和/或附加分箱运算的步骤包括：额外设置离群箱，使得具有离群值的连续特征被分到所述离群箱。

[0040] 可选地，在所述系统中，模型训练装置基于对数几率回归算法来训练特征池模型。

[0041] 可选地，在所述系统中，特征池模型的效果包括特征池模型的AUC。

[0042] 可选地，在所述系统中，所述原始测试数据集由获取的历史数据记录构成，其中，模型训练装置将获取的历史数据记录划分为多组历史数据记录以逐步地训练各个特征池模型，并且，模型训练装置还使用经过当前组历史数据记录训练后的特征池模型来针对下一组历史数据记录执行预测以得到与所述下一组历史数据记录相应的分组AUC，并综合各个分组AUC来得到特征池模型的AUC，其中，在得到与所述下一组历史数据记录相应的分组AUC之后，利用所述下一组历史数据记录来继续训练经过所述当前组历史数据记录训练后的特征池模型。

[0043] 可选地，在所述系统中，模型训练装置在使用经过当前组历史数据记录训练后的特征池模型来针对下一组历史数据记录执行预测时，当所述下一组历史数据记录包括缺少用于产生特征池模型所基于的至少一部分特征的属性信息的缺失历史数据记录时，基于以下处理之一来得到与所述下一组历史数据记录相应的分组AUC：仅利用所述下一组历史数据记录中除了缺失历史数据记录以外的其他历史数据记录的预测结果来计算分组AUC；利

用所述下一组历史数据记录的全部历史数据记录的预测结果来计算分组AUC，其中，将缺失历史数据记录的预测结果设置为默认值，所述默认值基于预测结果的取值范围来确定或基于获取的历史数据记录的标记分布来确定；将利用所述下一组历史数据记录中除了缺失历史数据记录以外的其他历史数据记录的预测结果计算出的AUC与所述其他历史数据记录在所述下一组历史数据记录中所占的比例相乘来得到分组AUC。

[0044] 可选地，在所述系统中，模型训练装置在基于对数几率回归算法来训练特征池模型时，针对连续特征设置的正则项不同于针对非连续特征设置的正则项。

[0045] 可选地，所述系统还包括：显示装置，其中，模型训练装置还控制显示装置向用户提供用于配置特征池模型的以下项目之中的至少一个项目的界面：特征池模型所基于的至少一部分特征、特征池模型的算法种类、特征池模型的算法参数、离散化运算的运算种类、离散化运算的运算参数，并且，模型训练装置根据用户通过所述界面配置的项目来分别训练特征池模型。

[0046] 可选地，在所述系统中，模型训练装置响应于用户关于确定特征重要性的指示来控制显示装置向用户提供所述界面。

[0047] 可选地，在所述系统中，显示装置还以图形化方式向用户展示确定的各个特征的重要性。

[0048] 可选地，在所述系统中，显示装置按照特征的重要性的顺序来展示各个特征，并且/或者，对所述各个特征之中的一部分特征进行突出显示，其中，所述一部分特征包括与高重要性对应的重要特征、与低重要性对应的不重要特征和/或与异常重要性对应的异常特征。

[0049] 根据本发明的另一示例性实施例，提供一种确定机器学习样本的各个特征的重要性的计算装置，包括存储部件和处理器，存储部件中存储有计算机可执行指令集合，当所述计算机可执行指令集合被所述处理器执行时，执行下述步骤：(A)获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息；(B)利用获取的历史数据记录，训练至少一个特征池模型，其中，特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型；(C)获取所述至少一个特征池模型的效果，并根据获取的所述至少一个特征池模型的效果来确定所述各个特征的重要性，其中，在步骤(B)中，通过对所述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。

[0050] 可选地，在所述计算装置中，在步骤(C)中，根据特征池模型在原始测试数据集和变换测试数据集上的效果之间的差异来确定所述特征池模型所基于的相应特征的重要性，其中，变换测试数据集是指通过对原始测试数据集中的其重要性待确定的目标特征的取值替换为以下项之一而获得的数据集：零值、随机数值、通过将目标特征的原始取值扰乱顺序后得到的值。

[0051] 可选地，在所述计算装置中，所述至少一个特征池模型包括一个全部特征模型，其中，全部特征模型是指基于所述各个特征之中的全部特征来提供关于机器学习问题的预测结果的机器学习模型。

[0052] 可选地，在所述计算装置中，所述至少一个特征池模型包括多个基于不同特征组来提供关于机器学习问题的预测结果的机器学习模型，其中，在步骤(C)中，根据所述至少

一个特征池模型在原始测试数据集上的效果之间的差异来确定所述各个特征的重要性。

[0053] 可选地，在所述计算装置中，所述至少一个特征池模型包括一个或多个主特征池模型以及分别与每个主特征池模型相应的至少一个子特征池模型，其中，子特征池模型是指基于与其相应的主特征池模型所基于的特征之中除了其重要性待确定的目标特征之外的剩余特征来提供关于机器学习问题的预测结果的机器学习模型，其中，在步骤(C)中，根据主特征池模型和与其相应的各个子特征池模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

[0054] 可选地，在所述计算装置中，所述至少一个特征池模型包括多个单特征模型，其中，单特征模型是指基于所述各个特征之中的其重要性待确定的目标特征来提供关于机器学习问题的预测结果的机器学习模型，其中，在步骤(C)中，根据单特征模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

[0055] 可选地，在所述计算装置中，所述离散化运算包括基本分箱运算和至少一个附加运算。

[0056] 可选地，在所述计算装置中，所述至少一个附加运算包括以下种类的运算之中的至少一种运算：对数运算、指数运算、绝对值运算、高斯变换运算。

[0057] 可选地，在所述计算装置中，所述至少一个附加运算包括与基本分箱运算分箱方式相同但分箱参数不同的附加分箱运算；或者，所述至少一个附加运算包括与基本分箱运算分箱方式不同的附加分箱运算。

[0058] 可选地，在所述计算装置中，基本分箱运算和附加分箱运算分别对应于不同宽度的等宽分箱运算或不同深度的等深分箱。

[0059] 可选地，在所述计算装置中，所述不同宽度或不同深度在数值上构成等比数列或等差数列。

[0060] 可选地，在所述计算装置中，执行基本分箱运算和/或附加分箱运算的步骤包括：额外设置离群箱，使得具有离群值的连续特征被分到所述离群箱。

[0061] 可选地，在所述计算装置中，在步骤(B)中，基于对数几率回归算法来训练特征池模型。

[0062] 可选地，在所述计算装置中，特征池模型的效果包括特征池模型的AUC。

[0063] 可选地，在所述计算装置中，所述原始测试数据集由获取的历史数据记录构成，其中，在步骤(B)中，将获取的历史数据记录划分为多组历史数据记录以逐步地训练各个特征池模型，并且，步骤(B)还包括：使用经过当前组历史数据记录训练后的特征池模型来针对下一组历史数据记录执行预测以得到与所述下一组历史数据记录相应的分组AUC，并综合各个分组AUC来得到特征池模型的AUC，其中，在得到与所述下一组历史数据记录相应的分组AUC之后，利用所述下一组历史数据记录来继续训练经过所述当前组历史数据记录训练后的特征池模型。

[0064] 可选地，在所述计算装置中，在步骤(B)中，在使用经过当前组历史数据记录训练后的特征池模型来针对下一组历史数据记录执行预测时，当所述下一组历史数据记录包括缺少用于产生特征池模型所基于的至少一部分特征的属性信息的缺失历史数据记录时，基于以下处理之一来得到与所述下一组历史数据记录相应的分组AUC：仅利用所述下一组历史数据记录中除了缺失历史数据记录以外的其他历史数据记录的预测结果来计算分组

AUC;利用所述下一组历史数据记录的全部历史数据记录的预测结果来计算分组AUC,其中,将缺失历史数据记录的预测结果设置为默认值,所述默认值基于预测结果的取值范围来确定或基于获取的历史数据记录的标记分布来确定;将利用所述下一组历史数据记录中除了缺失历史数据记录以外的其他历史数据记录的预测结果计算出的AUC与所述其他历史数据记录在所述下一组历史数据记录中所占的比例相乘来得到分组AUC。

[0065] 可选地,在所述计算装置中,在步骤(B)中,在基于对数几率回归算法来训练特征池模型时,针对连续特征设置的正则项不同于针对非连续特征设置的正则项。

[0066] 可选地,在所述计算装置中,步骤(B)还包括:向用户提供用于配置特征池模型的以下项目之中的至少一个项目的界面:特征池模型所基于的至少一部分特征、特征池模型的算法种类、特征池模型的算法参数、离散化运算的运算种类、离散化运算的运算参数,并且,在步骤(B)中,根据用户通过所述界面配置的项目来分别训练特征池模型。

[0067] 可选地,在所述计算装置中,在步骤(B)中,响应于用户关于确定特征重要性的指示来向用户提供所述界面。

[0068] 可选地,在所述计算装置中,当所述计算机可执行指令集合被所述处理器执行时,还执行下述步骤:(D)以图形化方式向用户展示确定的各个特征的重要性。

[0069] 可选地,在所述计算装置中,在步骤(D)中,按照特征的重要性的顺序来展示各个特征,并且/或者,对所述各个特征之中的一部分特征进行突出显示,其中,所述一部分特征包括与高重要性对应的重要特征、与低重要性对应的不重要特征和/或与异常重要性对应的异常特征。

[0070] 在根据本发明示例性实施例的确定机器学习样本的特征重要性的方法及系统中,利用以机器学习样本的至少一部分特征为基础的特征池模型的效果来相应地确定各特征的重要性,其中,在训练特征池模型时,所述至少一部分特征之中的连续特征需经过离散化处理,这样,可有效地通过特征池模型的效果来反映相关特征的重要程度,进而有效地得出各特征的重要性。

附图说明

[0071] 从下面结合附图对本发明实施例的详细描述中,本发明的这些和/或其他方面和优点将变得更加清楚并更容易理解,其中:

[0072] 图1示出根据本发明示例性实施例的确定机器学习样本的特征重要性的系统的框图;

[0073] 图2示出根据本发明示例性实施例的确定机器学习样本的特征重要性的方法的流程图;

[0074] 图3示出根据本发明另一示例性实施例的确定机器学习样本的特征重要性的方法的流程图;

[0075] 图4示出根据本发明示例性实施例的特征重要性展示界面的示例;以及

[0076] 图5示出根据本发明另一示例性实施例的特征重要性展示界面的示例。

具体实施方式

[0077] 为了使本领域技术人员更好地理解本发明,下面结合附图和具体实施方式对本发

明的示例性实施例作进一步详细说明。

[0078] 在本发明的示例性实施例中，通过以下方式来确定特征重要性：基于机器学习样本的至少一部分特征来训练特征池模型，其中，连续特征需经过离散化处理。在此基础上，基于特征池模型的预测效果来衡量各个特征的重要性。

[0079] 这里，机器学习是人工智能研究发展到一定阶段的必然产物，其致力于通过计算的手段，利用经验来改善系统自身的性能。在计算机系统中，“经验”通常以“数据”形式存在，通过机器学习算法，可从数据中产生“模型”，也就是说，将经验数据提供给机器学习算法，就能基于这些经验数据产生模型，在面对新的情况时，模型会提供相应的判断，即，预测结果。不论是训练机器学习模型，还是利用训练好的机器学习模型进行预测，数据都需要转换为包括各种特征的机器学习样本。机器学习可被实现为“有监督学习”、“无监督学习”或“半监督学习”的形式，应注意，本发明对具体的机器学习算法并不进行特定限制。此外，还应注意，在训练和应用模型的过程中，还可结合统计算法等其他手段。

[0080] 图1示出根据本发明示例性实施例的确定机器学习样本的特征重要性的系统的框图。具体说来，所述特征重要性确定系统利用基于至少一部分特征的特征池模型的预测效果来衡量各个相应特征的重要性，其中，特征池模型所基于的至少一部分原始连续特征需经过离散化处理。通过上述方式，可更加有效地确定各个特征(特别是连续特征)的重要性。

[0081] 图1所示的系统可全部通过计算机程序以软件方式来实现，也可由专门的硬件装置来实现，还可通过软硬件结合的方式来实现。相应地，组成图1所示的系统的各个装置可以是仅依靠计算机程序来实现相应功能的虚拟模块，也可以是依靠硬件结构来实现所述功能的通用或专用器件，还可以是运行有相应计算机程序的处理器等。利用所述系统，能够确定出机器学习样本的各个特征的重要性，这些重要性信息有助于进行模型训练和/或模型解释。

[0082] 如图1所示，数据记录获取装置100用于获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息。

[0083] 上述历史数据记录可以是在线产生的数据、预先生成并存储的数据、也可以是通过输入装置或传输媒介而从外部装置接收的数据，例如，可以是云端从客户端接收的数据或者客户端从云端接收的数据。这些数据可涉及个人、企业或组织的信息，例如，身份、学历、职业、资产、联系方式、负债、收入、盈利、纳税等信息。或者，这些数据也可涉及业务相关项目的信息，例如，关于合同的交易额、交易双方、标的物、交易地点等信息。应注意，本发明的示例性实施例中提到的属性信息内容可涉及任何对象或事务在某方面的表现或性质，而不限于对个人、物体、组织、单位、机构、项目、事件等进行限定或描述。

[0084] 数据记录获取装置100可获取不同来源的结构化或非结构化数据，例如，文本数据或数值数据等。获取的历史数据记录可用于形成机器学习样本，参与机器学习模型的训练和/或测试。这些数据可来源于期望应用机器学习的实体内部，例如，来源于期望应用机器学习的银行、企业、学校等；这些数据也可来源于上述实体以外，例如，来源于数据提供商、互联网(例如，社交网站)、移动运营商、APP运营商、快递公司、信用机构等。可选地，上述内部数据和外部数据可组合使用，以形成携带更多信息的机器学习样本，从而更便于发掘出重要性较高的特征。

[0085] 上述数据可通过输入装置输入到数据记录获取装置100，或者由数据记录获取装置100根据已有的数据来自动生成，或者可由数据记录获取装置100从网络上(例如，网络上的存储介质(例如，数据仓库))获得，此外，诸如服务器的中间数据交换装置可有助于数据记录获取装置100从外部数据源获取相应的数据。这里，获取的数据可被数据记录获取装置100中的文本分析模块等数据转换模块转换为容易处理的格式。也就是说，数据记录获取装置100可以是具有接收并处理数据记录的能力的装置，也可以仅仅是提供已经准备好的数据记录的装置。应注意，数据记录获取装置100可被配置为由软件、硬件和/或固件组成的各个模块，这些模块中的某些模块或全部模块可被集成为一体或共同协作以完成特定功能。

[0086] 模型训练装置200用于利用获取的历史数据记录，训练至少一个特征池模型，其中，特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型，其中，模型训练装置200通过对所述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。

[0087] 这里，特征池模型被设计为基于机器学习样本的至少一部分特征，相应地，模型训练装置200可基于历史数据记录来产生特征池模型的训练样本。具体说来，假设历史数据记录具有属性信息$\{p_1,p_2,\cdots,p_m\}$和相应的标记(其中，m是正整数)，基于这些属性信息和标记，可产生与机器学习问题相应的机器学习样本，这些机器学习样本将应用于针对机器学习问题的模型训练和/或测试。具体说来，上述机器学习样本的特征部分可表示为$\{f_1,f_2,\cdots,f_n\}$(其中，n是正整数)，而本发明的示例性实施例旨在确定特征部分$\{f_1,f_2,\cdots,f_n\}$之中各个特征的重要程度。为此，模型训练装置200需训练出基于至少一部分特征来提供关于机器学习问题的预测结果的特征池模型，这里，模型训练装置200可从$\{f_1,f_2,\cdots,f_n\}$之中选择至少一部分特征作为特征池模型的训练样本的特征，并将相应历史数据记录的标记作为所述训练样本的标记。根据本发明的示例性实施例，所选择的至少一部分特征之中的部分或全部连续特征需经过离散化处理。这里，模型训练装置200可训练一个或多个特征池模型，其中，可基于相同特征池模型(所述相同特征池模型可基于机器学习样本的全部特征或一部分特征)在原始测试数据集与变换测试数据集上的预测效果差异来综合得出相应特征的重要性，其中，通过对原始测试数据集中的某些目标特征的取值进行变换来获得变换测试数据集，这样，预测效果差异即可反映出目标特征的预测作用，即，重要性；或者，可基于不同特征池模型在相同测试数据集(即，原始测试数据集)上的预测效果差异来综合得出相应特征的重要性，这里，不同特征池模型可被设计为基于不同的特征组合，这样，预测效果差异即可反映出不同特征各自的预测作用，即，重要性；特别地，可分别针对机器学习样本的每个特征来训练出单特征模型，相应地，单特征模型的预测效果即可代表其所依据的特征的重要性。应注意，上述两种衡量特征重要性的方式可单独使用，也可结合使用。

[0088] 如上所述，根据本发明的示例性实施例，在训练特征池模型时，模型训练装置200可通过对至少一个连续特征执行离散化运算来训练特征池模型，这里，模型训练装置200可采用任何适当的离散化方式对连续特征进行处理，以便基于离散化后的连续特征(或连同其他特征)所训练出的特征池模型能够更好地反映各个特征的重要程度。

[0089] 这里，作为示例，所述离散化运算可包括基本分箱(binning)运算和至少一个附加运算，相应地，模型训练装置200可在训练特征池模型时，针对特征池模型所依据的某些连续特征之中的每一个连续特征，分别执行基本分箱运算和至少一个附加运算，以产生与各

连续特征对应的基本分箱特征和至少一个附加特征。

[0090] 这里，在机器学习样本的特征之中，会存在基于数据记录的至少一部分属性信息所产生的连续特征，这里，连续特征是与离散特征(例如，类别特征)相对的一种特征，其取值可以是具有一定连续性的数值，例如，距离、年龄、金额等。相对地，作为示例，离散特征的取值不具有连续性，例如，可以是“来自北京”、“来自上海”或“来自天津”、“性别为男”、“性别为女”等无序分类的特征。

[0091] 举例说来，历史数据记录中的某种连续值属性可直接作为机器学习样本中的对应连续特征，例如，可将距离、年龄、金额等属性直接作为相应的连续特征。此外，也可通过对历史数据记录中的某些属性(例如，连续属性和/或离散属性)进行处理，以得到相应的连续特征，例如，将身高与体重的比值作为相应的连续特征。

[0092] 应注意，除了将进行基本分箱运算和附加运算的连续特征之外，特征池模型的训练样本还可包括依据机器学习样本所包括的其他连续特征和/或离散特征，其中，所述其他连续特征可在不经过离散化运算的情况下参与特征池模型的训练。

[0093] 可以看出，根据本发明的示例性实施例，对于将进行基本分箱运算的每一个连续特征，还可额外执行至少一个附加运算，从而能够同时获得多个从不同的角度、尺度/层面来刻画原始数据记录的某些属性的特征。

[0094] 这里，分箱运算是指将连续特征进行离散化的一种特定方式，即，将连续特征的值域划分为多个区间(即，多个箱子)，并基于划分的箱子来确定相应的分箱特征值。分箱运算大体上可划分为有监督分箱和无监督分箱，这两种类型各自包括一些具体的分箱方式，例如，有监督分箱包括最小熵分箱、最小描述长度分箱等，而无监督分箱包括等宽分箱、等深分箱、基于k均值聚类的分箱等。在每种分箱方式下，可设置相应的分箱参数，例如，宽度、深度等。应注意，根据本发明的示例性实施例，由模型训练装置200执行的分箱运算不限制分箱方式的种类，也不限制分箱运算的参数，并且，相应产生的分箱特征的具体表示方式也不受限制。

[0095] 除了执行基本分箱运算之外，模型训练装置200还可对所述连续特征执行至少一个附加运算，这里，附加运算可以是任意函数运算，这些函数运算可产生连续特征或离散特征，例如，附加运算可以是对数运算、指数运算、绝对值运算等。特别地，附加运算也可以是分箱运算(称为“附加分箱运算”)，这里的附加分箱运算与基本分箱运算在分箱方式和/或分箱参数方面存在差异。由此可见，所述至少一个附加运算可以是相同或不同种类的运算各自在相同或不同运算参数(例如，指数运算中的指数、对数运算中的底数、分箱运算中的深度、分箱运算中的宽度等)下的运算，这里，所述附加运算可以是以对数运算、指数运算、绝对值运算等为主体的表达式运算，也可以是多种运算的组合。

[0096] 通过上述方式，模型训练装置200可将至少一部分连续特征之中的每一个分别转换为基本分箱特征以及相应的至少一个附加特征，从而提升了用于特征池模型的机器学习素材的有效性，为后续的特征重要性确定提供了较好的基础。

[0097] 接下来，模型训练装置200可产生至少包括所产生的基本分箱特征和至少一个附加特征的训练样本，用于训练相应的特征池模型。这里，在所述训练样本中，除了由模型训练装置200产生的基本分箱特征和附加特征之外，还可包括任意的其他特征，其中，所述其他特征可以是属于应基于历史数据记录产生的机器学习样本中的特征。

[0098] 模型训练装置200可基于上述训练样本来训练特征池模型。这里，模型训练装置200可利用适当的机器学习算法(例如，对数几率回归)，从训练样本学习出适当的特征池模型。

[0099] 重要性确定装置300用于获取所训练出的至少一个特征池模型的效果，并根据获取的所述至少一个特征池模型的效果来确定所述各个特征的重要性。这里，重要性确定装置300可通过将训练出的特征池模型应用于相应的测试数据集来获取特征池模型的效果，也可从与其连接的其他方接收特征池模型的效果。

[0100] 具体说来，特征池模型在测试集上的表现可作为该特征池模型的预测效果，而这一预测效果可用于衡量特征池模型所基于的特征组的预测能力。通过衡量不同特征池模型在原始测试数据集上的效果差异或者相同特征池模型在不同测试特征上的效果差异，可综合得出机器学习样本的各特征的重要性。

[0101] 这里，作为示例，特征池模型的效果可包括特征池模型的AUC(ROC(受试者工作特征，Receiver Operating Characteristic)曲线下的面积，Area Under ROC Curve)。

[0102] 例如，假设某特征池模型所依据的特征为机器学习样本的特征部分 $\{f_1,f_2,\cdots,f_n\}$ 之中的三个特征 $\{f_1,f_3,f_5\}$，并且，其中的连续特征 f_1 在特征池模型的训练样本中是经过离散化处理的，相应地，该特征池模型在测试数据集上的AUC可反映特征组合 $\{f_1,f_3,f_5\}$ 的预测能力。此外，假设还有另一特征池模型所依据的两个特征为 $\{f_1,f_3\}$，同样地，连续特征 f_1 经过了离散化处理，相应地，该特征池模型在测试数据集上的AUC可反映特征组合 $\{f_1,f_3\}$ 的预测能力。在此基础上，上述两个AUC之间的差值可用于反映特征 f_5 的重要性。

[0103] 又例如，假设某特征池模型所依据的特征为机器学习样本的特征部分 $\{f_1,f_2,\cdots,f_n\}$ 之中的三个特征 $\{f_1,f_3,f_5\}$，并且，其中的连续特征 f_1 在特征池模型的训练样本中是经过离散化处理的，相应地，该特征池模型在原始测试数据集上的AUC可反映特征组合 $\{f_1,f_3,f_5\}$ 的预测能力。这里，为了确定目标特征 f_5 的重要性，可通过对原始测试数据集所包括的各个测试样本中的特征 f_5 的取值进行处理来得到变换测试数据集，并进而获得特征池模型在变换测试数据集上的AUC。在此基础上，上述两个AUC之间的差值可用于反映目标特征 f_5 的重要性。作为示例，在变换处理中，可将各原始测试样本中的特征 f_5 的取值替换为零值、随机数值、或通过将特征 f_5 的原始取值扰乱顺序后得到的值。

[0104] 应理解，上述各装置可被分别配置为执行特定功能的软件、硬件、固件或上述项的任意组合。例如，这些装置可对应于专用的集成电路，也可对应于纯粹的软件代码，还可对应于软件与硬件相结合的单元或模块。此外，这些装置所实现的一个或多个功能也可由物理实体设备(例如，处理器、客户端或服务器等)中的组件来统一执行。

[0105] 以下参照图2来描述根据本发明示例性实施例的确定机器学习样本的特征重要性的方法的流程图。这里，作为示例，图2所示的方法可由图1所示的特征重要性确定系统来执行，也可完全通过计算机程序以软件方式实现，还可通过特定配置的计算装置来执行图2所示的方法。为了描述方便，假设图2所示的方法由图1所示的特征重要性确定系统来执行。

[0106] 如图所示，在步骤S100中，由数据记录获取装置100获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息。

[0107] 这里，历史数据记录是关于期望预测的机器学习问题的真实记录，其包括属性信

息和标记两部分，这样的历史数据记录将用于形成机器学习样本，作为机器学习的素材，而本发明的示例性实施例旨在确定形成的机器学习样本中，各个特征的重要程度。

[0108] 具体说来，作为示例，数据记录获取装置100可通过手动、半自动或全自动的方式来采集历史数据，或对采集的原始历史数据进行处理，使得处理后的历史数据记录具有适当的格式或形式。作为示例，数据记录获取装置100可批量地采集历史数据。

[0109] 这里，数据记录获取装置100可通过输入装置(例如，工作站)接收用户手动输入的历史数据记录。此外，数据记录获取装置100可通过全自动的方式从数据源系统地取出历史数据记录，例如，通过以软件、固件、硬件或其组合实现的定时器机制来系统地请求数据源并从响应中得到所请求的历史数据。所述数据源可包括一个或多个数据库或其他服务器。可经由内部网络和/或外部网络来实现全自动获取数据的方式，其中可包括通过互联网来传送加密的数据。在服务器、数据库、网络等被配置为彼此通信的情况下，可在没有人工干预的情况下自动进行数据采集，但应注意，在这种方式下仍旧可存在一定的用户输入操作。半自动方式介于手动方式与全自动方式之间。半自动方式与全自动方式的区别在于由用户激活的触发机制代替了例如定时器机制。在这种情况下，在接收到特定的用户输入的情况下，才产生提取数据的请求。每次获取数据时，优选地，可将捕获的历史数据存储在非易失性存储器中。作为示例，可利用数据仓库来存储在获取期间采集的原始数据以及处理后的数据。

[0110] 上述获取的历史数据记录可来源于相同或不同的数据源，也就是说，每条历史数据记录也可以是不同历史数据记录的拼接结果。例如，除了获取客户向银行申请开通信用卡时填写的信息数据记录(其包括收入、学历、职务、资产情况等属性信息字段)之外，作为示例，数据记录获取装置100可还获取该客户在该银行的其他数据记录，例如，贷款记录、日常交易数据等，这些获取的数据记录可连同关于该客户是否为欺诈客户的标记拼接为完整的历史数据记录。此外，数据记录获取装置100还可获取来源于其他私有源或公共源的数据，例如，来源于数据提供商的数据、来源于互联网(例如，社交网站)的数据、来源于移动运营商的数据、来源于APP运营商的数据、来源于快递公司的数据、来源于信用机构的数据等等。

[0111] 可选地，数据记录获取装置100可借助硬件集群(诸如Hadoop集群、Spark集群等)对采集到的数据进行存储和/或处理，例如，存储、分类和其他离线操作。此外，数据记录获取装置100也可对采集的数据进行在线的流处理。

[0112] 作为示例，数据记录获取装置100中可包括文本分析模块等数据转换模块，相应地，在步骤S100中，数据记录获取装置100可将文本等非结构化数据转换为更易于使用的结构化数据以在后续进行进一步的处理或引用。基于文本的数据可包括电子邮件、文档、网页、图形、电子数据表、呼叫中心日志、交易报告等。

[0113] 接下来，在步骤S200中，由模型训练装置200利用获取的历史数据记录，训练至少一个特征池模型，其中，特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型，其中，通过对所述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。

[0114] 这里，模型训练装置200可针对所述至少一个连续特征分别执行任何适当的离散化运算，作为示例，模型训练装置200可执行基本分箱运算和至少一个附加运算，以产生与

各连续特征分别对应的基本分箱特征和至少一个附加特征，产生的基本分箱特征和至少一个附加特征作为离散化后的特征，可构成特征池模型的训练样本的至少一部分特征。

[0115] 如上所述，连续特征作为机器学习样本中的特征，其可产生自历史数据记录的至少一部分属性信息，例如，历史数据记录的距离、年龄和金额等连续取值的属性信息可直接作为连续特征，又例如，可通过对历史数据记录的某些属性信息进行进一步的处理来获得连续特征，比如，可将身高与体重的比值作为连续特征。

[0116] 在获得了连续特征之后，可由模型训练装置200对获得的连续特征执行基本分箱运算，这里，模型训练装置200可按照各种分箱方式和/或分箱参数来执行基本分箱运算。

[0117] 以无监督下的等宽分箱为例，假设连续特征的取值区间为[0,100]，相应的分箱参数(即，宽度)为50，则可分出2个箱子，在这种情况下，取值为61.5的连续特征对应于第2个箱子，如果这两个箱子的标号为0和1，则所述连续特征对应的箱子标号为1。或者，假设分箱宽度为10，则可分出10个箱子，在这种情况下，取值为61.5的连续特征对应于第7个箱子，如果这十个箱子的标号为0到9，则所述连续特征对应的箱子标号为6。或者，假设分箱宽度为2，则可分出50个箱子，在这种情况下，取值为61.5的连续特征对应于第31个箱子，如果这五十个箱子的标号为0到49，则所述连续特征对应的箱子标号为30。作为示例，可通过在线计算的方式来确定具体连续特征的箱子标号并得到相应的特征值，而不需要采用查找映射表的方式，从而节省存储空间开销。

[0118] 在将连续特征映射到多个箱子之后，对应的特征值可以为自定义的任何值。也就是说，执行基本分箱运算以产生与连续特征对应的多维度的基本分箱特征，其中，作为示例，每个维度可指示对应的箱子中是否被分到了相应的连续特征，例如，以“1”来表示连续特征被分到了相应的箱子，而以“0”来表示连续特征没有被分到相应的箱子，相应地，在上述示例中，假设分出了10个箱子，则基本分箱特征可以是10个维度的特征，与取值为61.5的连续特征对应的基本分箱特征可表示为[0,0,0,0,0,0,1,0,0,0]。或者，每个维度可指示对应的箱子中被分到的相应的连续特征的特征值，相应地，在上述示例中，与取值为61.5的连续特征对应的基本分箱特征可表示为[0,0,0,0,0,0,61.5,0,0,0]；或者，每个维度指示对应的箱子中被分到的所有连续特征的特征值的平均值；或者，每个维度指示对应的箱子中被分到的所有连续特征的特征值的中间值；或者，每个维度指示对应的箱子中被分到的所有连续特征的特征值的边界值，这里的边界值可以是上边界值或下边界值。

[0119] 除此之外，还可对基本分箱特征的取值进行归一化处理，以便于执行运算。假设将进行离散化运算的第i个连续特征的第j个值为x_{ij}，其分箱特征可表示为(BinID,x'_{ij})，其中，BinID指示连续特征被分到的箱子的标号，该标号的取值范围为0、1、…、B-1，其中，B为箱子的总数，x'_{ij}为x_{ij}的归一化值，上述特征(BinID,x'_{ij})表示基本分箱特征中与标号为BinID的箱子对应的维度的特征取值为x'_{ij}，其余维度的特征取值为0。

[0120] 其中，x'_{ij}可如下式表示：

[0121] $$x'_{ij} = \left(x_{ij} - \min_i\right) \times \frac{B}{\max_i - \min_i} - \mathrm{BinID},$$

[0122] 其中，$\max_i$为第i个连续特征的最大值，$\min_i$为第i个连续特征的最小值，并且，

[0123] $\mathrm{BinID} = \left\lfloor \left(x_{ij} - \min_i\right) \times \frac{B}{\max_i - \min_i} \right\rfloor$，其中，[ ]为向下取整运算符号。

[0124] 以无监督下的等宽分箱为例，假设连续特征的取值区间为[0,100]，在分箱宽度为50的情况下，按照上述计算式，取值为61.5的连续特征可对应于基本分箱特征(1,0.23)，而在分箱宽度为10的情况下，按照上述计算式，取值为61.5的连续特征可对应于基本分箱特征(6,0.15)。

[0125] 这里，为了获得上述特征(BinID,x'_{ij})，在步骤S200中，模型训练装置200可按照上述计算式，针对每一个x_{ij}值进行BinID和x'_{ij}的运算，或者，模型训练装置200也可预先产生关于各个BinID的取值范围的映射表，通过查找该数据表来获得与连续特征相应的BinID。

[0126] 此外，作为示例，在执行基本分箱运算前，还可以通过去除连续特征中的离群点来减少历史数据记录中的噪音。通过这种方式，能进一步提高利用分箱特征来确定特征重要性的有效性。

[0127] 具体说来，可额外设置离群箱，使得具有离群值的连续特征被分到所述离群箱。举例说来，对于取值区间为[0,1000]的连续特征，可选取一定数量的样本进行预分箱，例如，先按照分箱宽度为10来进行等宽分箱，然后记录每个箱子内的样本数量，对于样本数量较少(例如，少于阈值)的箱子，可以将它们合并为至少一个离群箱。作为示例，如果位于两端的箱内样本数量较少，则可将样本较少的箱子合并为离群箱，而将剩余的箱子保留，假设0-10号箱子中的样本数量较少，则可将0-10号箱子合并为离群箱，从而将取值为[0,100]的连续特征统一划分到离群箱。

[0128] 除了执行上述基本分箱运算以外，在步骤S200中，模型训练装置200还针对所述被执行基本分箱运算的连续特征，执行至少一个不同于基本分箱运算的附加运算以获得相应的至少一个附加特征。

[0129] 这里，所述附加运算可以是任意的函数运算，这些函数运算可具有相应的运算参数，并且，针对单个连续特征执行的附加运算可以是一个或多个运算，所述多个运算可以是不同种类的运算，也可以是相同种类但不同运算参数的运算。

[0130] 特别地，附加运算也可指示分箱运算，这里，类似于基本分箱特征，通过附加分箱运算产生的附加分箱特征也可以是多维度的特征，其中，每个维度指示对应的箱子中是否被分到了相应的连续特征；或者，每个维度指示对应的箱子中被分到的相应的连续特征的特征值；或者，每个维度指示对应的箱子中被分到的所有连续特征的特征值的平均值；或者，每个维度指示对应的箱子中被分到的所有连续特征的特征值的中间值；或者，每个维度指示对应的箱子中被分到的所有连续特征的特征值的边界值。

[0131] 具体说来，所述至少一个附加运算可包括与基本分箱运算分箱方式相同但分箱参数不同的附加分箱运算；或者，所述至少一个附加运算可包括与基本分箱运算分箱方式不同的附加分箱运算。这里的分箱方式包括有监督分箱和/或无监督分箱下的各种分箱方式。例如，有监督分箱包括最小熵分箱、最小描述长度分箱等，而无监督分箱包括等宽分箱、等深分箱、基于k均值聚类的分箱等。

[0132] 作为示例，基本分箱运算和附加分箱运算可分别对应于不同宽度的等宽分箱运算。也就是说，基本分箱运算和附加分箱运算采用的分箱方式相同但划分的粒度不同，这使得产生的基本分箱特征和附加分箱特征能够更好地刻画原始历史数据记录的规律，从而更有利于确定各特征的重要性。特别地，基本分箱运算和附加分箱运算所采用的不同宽度可在数值上构成等比数列，例如，基本分箱运算可按照值2的宽度来进行等宽分箱，而附加分

箱运算可按照值4、值8、值16等的宽度来进行等宽分箱。或者，基本分箱运算和附加分箱运算所采用的不同宽度可在数值上构成等差数列，例如，基本分箱运算可按照值2的宽度来进行等宽分箱，而附加分箱运算可按照值4、值6、值8等的宽度来进行等宽分箱。

[0133] 作为另一示例，基本分箱运算和附加分箱运算可分别对应于不同深度的等深分箱运算。也就是说，基本分箱运算和附加分箱运算采用的分箱方式相同但划分的粒度不同，这使得产生的基本分箱特征和附加分箱特征能够更好地刻画原始历史数据记录的规律，从而更有利于确定各特征的重要性。特别地，基本分箱运算和附加分箱运算所采用的不同深度可在数值上构成等比数列，例如，基本分箱运算可按照值10的深度来进行等深分箱，而附加分箱运算可按照值100、值1000、值10000等的深度来进行等深分箱。或者，基本分箱运算和附加分箱运算所采用的不同深度可在数值上构成等差数列，例如，基本分箱运算可按照值10的深度来进行等深分箱，而附加分箱运算可按照值20、值30、值40等的深度来进行等深分箱。

[0134] 根据本发明的示例性实施例，附加运算还可包括非分箱运算，例如，所述至少一个附加运算包括以下种类的运算之中的至少一种运算各自在相同或不同运算参数下的运算：对数运算、指数运算、绝对值运算、高斯变换运算。应注意，这里的附加运算不受运算种类和运算参数的限制，可采用任何适当的算式形式，也就是说，附加运算既可以具有诸如平方运算这样的简单形式，也可以具有复杂的运算表达式，例如，对于第i个连续特征的第j个值x_{ij}，可按照下式对其执行附加运算以得到附加特征x''_{ij}：

[0135] $x''_{ij}=\mathrm{sign}(x_{ij})\times\log_2(1+|x_{ij}|)$，其中，sign为符号函数。

[0136] 除了上述基本分箱特征和附加特征之外，还可产生特征池模型的训练样本中包括的其他特征，这些特征可由模型训练装置200通过对历史数据记录的至少一部分属性信息进行诸如直接提取、离散化、字段组合、提取部分字段值、取整等各种特征处理而获得。

[0137] 接下来，由模型训练装置200产生包括上述特征连同相应的标记的特征池模型的训练样本。根据本发明的示例性实施例，可在分布式并行计算框架下于内存中执行上述处理，这里的分布式并行计算框架可具有分布式参数服务器。

[0138] 此外，作为示例，产生的训练样本可被直接用于特征池模型的训练处理。具体说来，产生所述训练样本的步骤可被视为特征池模型训练过程的一部分，相应地，训练样本不需要显式地保存到硬盘，这种处理方式与传统方式相比可明显地提高运行速度。

[0139] 接下来，可由模型训练装置200基于训练样本来训练特征池模型。这里，模型训练装置200可利用适当的机器学习算法(例如，对数几率回归)，从训练样本学习出适当的特征池模型。作为示例，在特征池模型的训练样本既包括连续特征也包括非连续特征的情况下，可分别针对连续特征和非连续特征设置不同的正则项，即，针对连续特征设置的正则项不同于针对非连续特征设置的正则项。

[0140] 在上述示例中，可训练出较为稳定且预测效果较好的特征池模型，以便于后续基于特征池模型的预测效果来有效地确定各特征的重要性。

[0141] 具体说来，在步骤S300中，由重要性确定装置300获取所训练出的至少一个特征池模型的效果，并根据获取的所述至少一个特征池模型的效果来确定机器学习样本的各个特征的重要性。

[0142] 这里，重要性确定装置300可通过将训练出的特征池模型应用于相应的测试数据

集来获取特征池模型的效果，也可从与其连接的其他方接收特征池模型的效果。

[0143] 作为示例，重要性确定装置300可根据特征池模型在原始测试数据集和变换测试数据集上的效果之间的差异来确定所述特征池模型所基于的相应特征的重要性，其中，变换测试数据集是指通过对原始测试数据集中的其重要性待确定的目标特征的取值替换为以下项之一而获得的数据集：零值、随机数值、通过将目标特征的原始取值扰乱顺序后得到的值。

[0144] 这里，每个特征池模型可基于机器学习样本的至少一个特征，相应地，可获得所述特征池模型在原始测试数据集上的预测效果。此外，可通过变换原始测试数据集上的目标特征的取值来获取所述特征池模型在变换测试数据集上的预测效果。上述两个预测效果的差异即可用来衡量目标特征的重要性。

[0145] 作为示例，所述至少一个特征池模型可包括一个全部特征模型，其中，全部特征模型是指基于机器学习样本的各个特征之中的全部特征来提供关于机器学习问题的预测结果的机器学习模型，具体说来，假设模型训练装置200在步骤S200中训练出一个全部特征模型，该全部特征模型被训练为基于机器学习样本的全部特征 $\{f_1, f_2, \cdots, f_n\}$ 来给出关于机器学习问题的预测结果。重要性确定装置300可获取该全部特征模型在原始测试数据集上的预测效果(例如，AUC_{all})，这里的原始测试数据集可产生自由数据记录获取装置100获取的其他历史数据记录。

[0146] 在这一示例中，为了确定 $\{f_1, f_2, \cdots, f_n\}$ 之中的任一目标特征 f_i 的重要性(其中，$1 \leqslant i \leqslant n$)，可相应地对原始测试数据集进行处理以得到针对目标特征 f_i 的变换测试数据集，例如，将原始测试数据集的各个测试样本中的特征 f_i 的取值替换为其他值，例如，零值、随机数值、或者将特征 f_i 的取值在各个测试样本之间打乱顺序之后获得的值。相应地，重要性确定装置300可获取上述全部特征模型在变换测试数据集上的测试效果(例如，AUC_i)。

[0147] 在分别获取了全部特征模型在原始测试数据集和变换测试数据集上的效果之后，重要性确定装置300可将两个效果之间的差异(即，$AUC_{all}-AUC_i$)作为衡量目标特征 f_i 的重要性的参考。

[0148] 以上示出了通过对原始测试数据集进行变换，从而借助同样的特征池模型来确定其所依据的各个特征的重要性的示例。然而，本发明的示例性实施例并不受限于此，可采用任何适当的方式来设计特征池模型的个数以及各个特征池模型所依据的特征组，只要这些特征池模型的预测效果可推断出各个特征的重要性即可。

[0149] 例如，由模型训练装置200在步骤S200中所训练出的至少一个特征池模型可包括多个基于不同特征组来提供关于机器学习问题的预测结果的机器学习模型，相应地，在步骤S300中，重要性确定装置300可根据所述至少一个特征池模型在原始测试数据集上的效果之间的差异来确定所述各个特征的重要性。

[0150] 这里，所述至少一个特征池模型包括一个或多个主特征池模型以及分别与每个主特征池模型相应的至少一个子特征池模型，其中，子特征池模型是指基于与其相应的主特征池模型所基于的特征之中除了其重要性待确定的目标特征之外的剩余特征来提供关于机器学习问题的预测结果的机器学习模型，相应地，重要性确定装置300可根据主特征池模型和与其相应的各个子特征池模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

[0151] 作为示例，所述至少一个特征池模型可包括一个作为主特征池模型的全部特征模型以及相应的至少一个子特征池模型，其中，全部特征模型是指基于机器学习样本的全部特征来提供关于机器学习问题的预测结果的机器学习模型，相应地，子特征池模型是指基于所述全部特征之中的除了其重要性待确定的目标特征之外的剩余特征来提供关于机器学习问题的预测结果的机器学习模型，相应地，在步骤S300中，重要性确定装置300可根据全部特征模型与各个子特征池模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

[0152] 具体说来，假设模型训练装置200在步骤S200中训练出一个全部特征模型，该全部特征模型被训练为基于机器学习样本的全部特征 $\{f_1, f_2, \cdots, f_n\}$ 来给出关于机器学习问题的预测结果。重要性确定装置300可获取该全部特征模型在原始测试数据集上的预测效果(例如，AUC_{all})，这里的原始测试数据集可产生自由数据记录获取装置100获取的其他历史数据记录。

[0153] 在这一示例中，为了确定 $\{f_1, f_2, \cdots, f_n\}$ 之中的任一目标特征 f_i 的重要性(其中，$1 \leqslant i \leqslant n$)，还可在步骤S200中额外确定相应的子特征池模型，该子特征池模型被训练为基于除了目标特征 f_i 的其他特征 $\{f_1, f_2, \cdots, f_{i-1}, f_{i+1}, \cdots, f_n\}$ 来给出关于机器学习问题的预测结果。相应地，重要性确定装置300可获取该子特征池模型在原始测试数据集上的预测效果(例如，AUC_i)。

[0154] 在分别获取了全部特征模型和各个子特征池模型在原始测试数据集上的效果之后，重要性确定装置300可将两个效果之间的差异(即，$AUC_{all}-AUC_i$)作为衡量所述特征 f_i 的重要性的参考。

[0155] 这里，应注意，上述全部特征模型仅作为示例，而非用于限制本发明示例性实施例的范围。实际上，在特征池模型中，可存在多个主特征池模型，每个主特征池模型具有各自的子特征池模型，也就是说，每个主特征池模型可基于机器学习样本的至少一部分特征，这里，不同主特征池模型之间可涉及或不涉及共同的特征。

[0156] 此外，作为可选方式，由模型训练装置200在步骤S200中所训练出的至少一个特征池模型可包括多个单特征模型，其中，单特征模型是指基于机器学习样本的各个特征之中的其重要性待确定的目标特征来提供关于机器学习问题的预测结果的机器学习模型，相应地，在步骤S300中，重要性确定装置300可根据各个单特征模型在原始测试数据集上的效果之间的差异来确定相应的目标特征的重要性。

[0157] 具体说来，假设模型训练装置200在步骤S200中训练出多个单特征模型，每个单特征模型被训练为基于机器学习样本的某个特征 $\{f_i\}$ 来给出关于机器学习问题的预测结果。这里，单特征模型的个数可与机器学习样本的特征个数相同。相应地，重要性确定装置300可获取各个单特征模型在相同测试数据集(例如，原始测试数据集)上的预测效果(例如，AUC_i)。这里，由于已针对连续特征进行过离散化处理(优选地，可执行过基本分箱运算和附加运算)，可确保单特征模型能够较为稳定地反映各个特征的预测能力，相应地，在分别获取了全部单特征模型在相同的测试数据集上的效果之后，重要性确定装置300可基于各个效果之间的差异来获取相应的各个特征之间的相对重要程度。

[0158] 以上参照图2示出了根据本发明示例性实施例的确定特征重要性的方法，然而，应理解，图2所示的方法并非用于限制本发明示例性实施例的具体实现方式，而只是提供了关

于本发明示例性实施例的基本构思的示例性说明，实际上，本领域技术人员可按照任何适当的方式通过对图2所示的方案进行变型和/或具体化来实施本发明的示例性实施例。举例来说，图2所示的流程图之中的各个步骤并非作为时序方面的任何限制，例如，步骤S200和步骤S300无需限定为严格的顺序执行，作为可选方式，可在训练特征池模型的过程中完成一部分模型测试操作以确定特征池模型的效果。

[0159] 具体说来，如上所述，根据本发明的示例性实施例，在步骤S200中，所训练的至少一个特征池模型可包括多个基于不同特征组来提供关于机器学习问题的预测结果的机器学习模型，并且，在步骤S300中，可根据所述至少一个特征池模型在原始测试数据集上的效果之间的差异来确定所述各个特征的重要性。

[0160] 这里，原始测试数据集可由获取的历史数据记录构成，相应地，在步骤S200中，将获取的历史数据记录划分为多组历史数据记录以逐步地训练各个特征池模型，并且，步骤S200还包括：使用经过当前组历史数据记录训练后的特征池模型来针对下一组历史数据记录执行预测以得到与所述下一组历史数据记录相应的分组AUC，并综合各个分组AUC来得到特征池模型的AUC，其中，在得到与所述下一组历史数据记录相应的分组AUC之后，可利用所述下一组历史数据记录来继续训练经过所述当前组历史数据记录训练后的特征池模型。

[0161] 图3示出根据本发明另一示例性实施例的确定机器学习样本的特征重要性的方法的流程图。同样地，为了描述方便，假设图3所示的方法由图1所示的特征重要性确定系统来执行。并且，作为示例，这里的特征池模型可以是基于对数几率回归算法的机器学习模型，而特征池模型的效果可由AUC来表示。

[0162] 参照图3，在步骤S100中，由数据记录获取装置100获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息。这里，为了简明，将不再赘述数据记录获取装置100获取历史数据记录的各种细节。

[0163] 接下来，在步骤S210中，由模型训练装置200将获取的历史数据记录划分为多组历史数据记录，这些划分出的多组历史数据记录将分批地逐步训练特征池模型。作为可选方式，所述训练过程可在线执行，在这种情况下，特征池模型的训练样本不需要显式地保存到硬盘。

[0164] 在步骤S220中，由模型训练装置200获取作为下一组历史数据记录的第k组历史数据记录，其中，k为正整数。根据本发明的示例性实施例，由于利用多组历史数据记录来分批地逐步训练各个特征池模型，因此，可理解：在获取第k组历史数据记录之前，已经根据之前的k-1批历史数据记录阶段性地训练出了各个特征池模型，这里，可将其中的特定特征池模型表示为LR_{k-1}。

[0165] 在步骤S230中，由模型训练装置200分别获取所训练出的一个或多个特征池模型在第k组历史数据记录的测试下所获取的相应分组AUC。以上述特定特征池模型LR_{k-1}为例，由模型训练装置200使用该特征池模型LR_{k-1}来针对第k组历史数据记录执行预测以得到与第k组历史数据记录相应的分组AUC，即，AUC_k。具体说来，为了将第k组历史数据记录用作测试数据集，需基于第k组历史数据记录之中的各条历史数据记录来生成测试样本，其中，测试样本的特征部分与特征池模型的训练样本的特征部分相一致，即，模型训练装置200可按照与训练样本类似的特征工程处理来得到测试样本的特征部分，同时舍弃历史数据记录的

标记，从而得到特征池模型的测试样本。接着，模型训练装置200将得到的测试样本输入特征池模型，以得到相应的预测结果。基于这些预测结果，模型训练装置200可获取所述特征池模型LR_{k-1}针对第k组历史数据记录的分组AUC_k。通过类似的方式，模型训练装置200可获取之前所训练出的所有特征池模型针对第k组历史数据记录的分组AUC，并保存这些分组AUC。

[0166] 实践中，有些历史数据记录中可能会缺少某些属性信息，而这些属性信息与特征池模型的特征相关，在这种情况下，为了更好地获取所述特征池模型的AUC，模型训练装置200可采取相应的应对处理。

[0167] 具体说来，在使用经过当前组历史数据记录训练后的特征池模型来针对下一组历史数据记录执行预测时，当所述下一组历史数据记录包括缺少用于产生特征池模型所基于的至少一部分特征的属性信息的缺失历史数据记录时，模型训练装置200可基于以下处理之一来得到与所述下一组历史数据记录相应的分组AUC：

[0168] 第一种情况：模型训练装置200可仅利用所述下一组历史数据记录中除了缺失历史数据记录以外的其他历史数据记录的预测结果来计算分组AUC。具体说来，假设第k组历史数据记录共包括1000条历史数据记录，其中，只有100条历史数据记录包括特征池模型的特征部分所基于的所有属性信息，即，有900条历史数据记录属于缺失历史数据记录。在这种情况下，模型训练装置200可仅利用所述100条具有完整相关属性信息的历史数据记录进行预测，并将基于预测结果获得的AUC作为分组AUC。

[0169] 第二种情况：模型训练装置200可利用所述下一组历史数据记录的全部历史数据记录的预测结果来计算分组AUC，其中，将缺失历史数据记录的预测结果设置为默认值，所述默认值基于预测结果的取值范围来确定或基于获取的历史数据记录的标记分布来确定。具体说来，假设第k组历史数据记录共包括1000条历史数据记录，其中，只有100条历史数据记录包括特征池模型的特征部分所基于的所有属性信息，即，有900条历史数据记录属于缺失历史数据记录。在这种情况下，模型训练装置200可将所述100条具有完整相关属性信息的历史数据记录输入特征池模型以进行预测，并将900条历史数据记录的预测结果设置为默认值，这里，作为示例，所述默认值可基于预测结果的取值范围来确定，例如，在预测结果的取值范围为[0,1]的情况下，可将所述默认值设置为中间值0.5；或者，所述默认值也可基于获取的历史数据记录的标记分布来确定，例如，假设在第k组历史数据记录所包括的1000条历史数据记录中，共有300条正样本(即，标记为1)，则可将所述默认值设置为正样本的概率，例如，0.3。在如上所述获得了全部1000条历史数据记录的相应预测结果时，模型训练装置200可将基于所述预测结果获得的AUC作为分组AUC。

[0170] 第三种情况：模型训练装置200可将利用所述下一组历史数据记录中除了缺失历史数据记录以外的其他历史数据记录的预测结果计算出的AUC与所述其他历史数据记录在所述下一组历史数据记录中所占的比例相乘来得到分组AUC。具体说来，假设第k组历史数据记录共包括1000条历史数据记录，其中，只有100条历史数据记录包括特征池模型的特征部分所基于的所有属性信息，即，有900条历史数据记录属于缺失历史数据记录。在这种情况下，模型训练装置200可将所述100条具有完整相关属性信息的历史数据记录输入特征池模型以进行预测，基于得到的预测结果来获取相应的AUC，接着，模型训练装置200可将获取的AUC乘以非缺失历史数据记录所占的比例(即，0.1)来确定最终的分组AUC。

[0171] 应注意，上述三种情况仅作为存在缺失历史数据记录时的示例性处理方式，而非用于限制本发明的示例性实施例。任何与上述三种方式相似或等同的方式也可应用于本发明的示例性实施例。

[0172] 在执行完特征池模型的测试之后，在步骤S240中，由模型训练装置200分别基于读取的第k组历史数据记录来继续训练截至到目前所训练出的一个或多个特征池模型。

[0173] 以上述特定特征池模型LR_{k-1}为例，在步骤S240中，由模型训练装置200使用第k组历史数据记录来继续进行模型训练以得到更新的特征池模型LR_k。具体说来，为了将第k组历史数据记录用作训练数据集，需基于第k组历史数据记录之中的各条历史数据记录来生成训练样本，即，模型训练装置200可按照相应的特征工程处理来得到训练样本的特征部分，同时将历史数据记录的标记作为训练样本的标记，从而得到特征池模型的训练样本。接着，模型训练装置200基于得到的训练样本继续训练特征池模型，以得到更新的特征池模型LR_k。通过类似的方式，模型训练装置200可利用第k组历史数据记录来更新之前所训练出的所有特征池模型。

[0174] 可以看出，根据本发明的示例性实施例，在分阶段训练特征池模型的过程中，可同时获取相应的分组AUC，这使得模型的训练和测试更为高效快速，实现了整个系统的优化。实际上，上述实例中得到的AUC与真实测试AUC的相关性很强(经测试，在特定数据集中，相关性可达到0.85以上)，因此，作为示例，可基于按照上述方式获得的分组AUC来确定特征池模型的各个特征的重要性。

[0175] 接下来，在步骤S250中，由模型训练装置200来确定获取的第k组历史数据记录是否是划分出的最后一组历史数据记录。如果在步骤S250中确定当前的第k组历史数据记录并非最后一组历史数据记录，则返回步骤S220以获取下一组划分的历史数据记录，即，第k+1组历史数据记录。相反，如果在步骤S250中确定当前的第k组历史数据记录为最后一组历史数据记录，则进行到步骤S310，在该步骤中，由重要性确定装置300基于所保存的各个特征池模型的分组AUC来确定机器学习样本的各个特征的重要性。

[0176] 具体说来，在步骤S310中，重要性确定装置300可将每个特征池模型的各个分组AUC进行综合，以得出代表相应特征池模型的性能的AUC。

[0177] 在获得了各个特征池模型的性能(即，AUC)之后，重要性确定装置300可将特征池模型的性能视作该特征池模型所涉及到的特征组(即，重要性待确定的机器学习样本中的各特征之中的至少一部分特征)的重要性参考，并通过综合各个特征池模型之间的性能差异来推算出每个目标特征的重要性或各个目标特征之间的重要性排序。

[0178] 同样地，应注意：图3所示的流程图也并非用于限制时序等处理上的细节，而仅用于作为示例来解释本发明的示例性实施例。作为示例，各个特征池模型的训练/测试是可以并行地和/或在线地执行的。

[0179] 根据本发明的示例性实施例，针对机器学习中使用的机器学习样本，可有效地确定其中所包括的各个特征的重要性程度，从而有助于更好地进行模型训练和/或模型解释。

[0180] 作为可选方式，图1所示的特征重要性确定系统可还包括显示装置(未示出)，相应地，在图2所示的步骤S200中，可由模型训练装置200控制显示装置向用户提供用于配置特征池模型的以下项目之中的至少一个项目的界面：特征池模型所基于的至少一部分特征、特征池模型的算法种类、特征池模型的算法参数、离散化运算的运算种类、离散化运算的运

算参数。此外，在该步骤中，模型训练装置200可根据用户通过所述界面配置的项目来分别训练特征池模型。这里，作为示例，在步骤S200中，可响应于用户关于确定特征重要性的指示来向用户提供所述界面。例如，在机器学习模型的训练过程中，为了确定相应的机器学习训练样本中各个特征的重要性情况，用户可在特征工程的过程中作出指示以期望获取各个特征的重要性。为此，根据本发明的示例性实施例，可在特征工程或建模流程的其他相关界面下向用户提供例如特征重要性算子的控件，当用户点击该控件时，即可向用户展示关于配置特征池模型的界面，在该界面中，可设置特征池模型的算法、特征、正则项等各个项目，特别是，还可设置关于对特征池模型的连续特征如何进行离散化的项目（例如，分箱运算的各种参数等）。例如，作为可选方式，可分别设置连续特征和非连续特征的正则项，还可分别设置不同连续特征所对应的正则项的不同权重。

[0181] 这里，所述显示装置可以是单纯的显示屏，在这种情况下，所述特征重要性确定系统还可包括便于用户通过所述界面来配置项目的输入装置（例如，键盘、鼠标、麦克风、摄像装置等）；或者，所述显示装置可以是具有触摸输入功能的触摸显示屏，在这种情况下，用户可直接通过该触摸屏来完成界面上的项目配置。

[0182] 此外，在根据本发明示例性实施例的特征重要性确定系统获取了机器学习样本的各个特征的重要性之后，还可通过图形化方式向用户展示所确定的各个特征的重要性信息。

[0183] 图4示出根据本发明示例性实施例的特征重要性展示界面的示例，在图4所示的界面中，展示了特征重要性分析报告，其中，列出了特征重要性排序以及其他的一些附加信息，作为示例，在点击或移动到某个特征的指示条时，还可额外显示关于该特征的样本信息或属性信息等。

[0184] 作为可选方式，可按照特征的重要性顺序来展示各个特征，并且/或者，对所述各个特征之中的一部分特征进行突出显示，其中，所述一部分特征包括与高重要性对应的重要特征、与低重要性对应的不重要特征和/或与异常重要性对应的异常特征。

[0185] 图5示出根据本发明另一示例性实施例的特征重要性展示界面的示例，在图5所示的界面中，不仅按照重要性的顺序示出了机器学习样本的各个特征，还对与异常重要性对应的异常特征进行了突出显示，可选地，进一步提供了出现该异常特征的可能原因，增强了用户交互体验。

[0186] 应理解：在现有的机器学习领域，多数情况下都需要依靠程序员编写代码来完成机器学习过程，即便是已经开发出一些诸如建模平台的软件系统，仍旧面临着难以惠及除了机器学习专家以外的业务人员的难题。然而，根据本发明的示例性实施例，能够有效地自动确定出机器学习样本中各个特征的重要性，使得应用机器学习的门槛有所降低。此外，根据本发明的示例性实施例，还可通过友好的交互方式向用户展示关于特征重要性的确定结果和/或关于确定方式的相关设置，进一步增强了机器学习平台的易用性，相应地，具备较高机器学习技术能力的用户可方便地设置和/或调整确定过程中的细节，而普通用户也可直观地了解到机器学习样本之中的重要特征、非重要特征和/或异常特征等。

[0187] 应注意，根据本发明示例性实施例的特征重要性系统可完全依赖计算机程序的运行来实现相应的功能，即，各个装置与计算机程序的功能架构中的各步骤相应，使得整个系统通过专门的软件包（例如，lib库）而被调用，以实现相应的功能。

[0188] 另一方面，特征重要性系统中的各个装置也可以通过硬件、软件、固件、中间件、微代码或其任意组合来实现。当以软件、固件、中间件或微代码实现时，用于执行相应操作的程序代码或者代码段可以存储在诸如存储介质的计算机可读介质中，使得处理器可通过读取并运行相应的程序代码或者代码段来执行相应的操作。

[0189] 这里，本发明的示例性实施例还可以实现为计算装置，该计算装置包括存储部件和处理器，存储部件中存储有计算机可执行指令集合，当所述计算机可执行指令集合被所述处理器执行时，执行上述特征重要性确定方法。

[0190] 具体说来，所述计算装置可以部署在服务器或客户端中，也可以部署在分布式网络环境中的节点装置上。此外，所述计算装置可以是PC计算机、平板装置、个人数字助理、智能手机、web应用或其他能够执行上述指令集合的装置。

[0191] 这里，所述计算装置并非必须是单个的计算装置，还可以是任何能够单独或联合执行上述指令(或指令集)的装置或电路的集合体。计算装置还可以是集成控制系统或系统管理器的一部分，或者可被配置为与本地或远程(例如，经由无线传输)以接口互联的便携式电子装置。

[0192] 在所述计算装置中，处理器可包括中央处理器(CPU)、图形处理器(GPU)、可编程逻辑装置、专用处理器系统、微控制器或微处理器。作为示例而非限制，处理器还可包括模拟处理器、数字处理器、微处理器、多核处理器、处理器阵列、网络处理器等。

[0193] 上述关于特征重要性确定方法中所描述的某些操作可通过软件方式来实现，某些操作可通过硬件方式来实现，此外，还可通过软硬件结合的方式来实现这些操作。

[0194] 处理器可运行存储在存储部件之一中的指令或代码，其中，所述存储部件还可以存储数据。指令和数据还可经由网络接口装置而通过网络被发送和接收，其中，所述网络接口装置可采用任何已知的传输协议。

[0195] 存储部件可与处理器集成为一体，例如，将RAM或闪存布置在集成电路微处理器等之内。此外，存储部件可包括独立的装置，诸如，外部盘驱动、存储阵列或任何数据库系统可使用的其他存储装置。存储部件和处理器可在操作上进行耦合，或者可例如通过I/O端口、网络连接等互相通信，使得处理器能够读取存储在存储部件中的文件。

[0196] 此外，所述计算装置还可包括视频显示器(诸如，液晶显示器)和用户交互接口(诸如，键盘、鼠标、触摸输入装置等)。计算装置的所有组件可经由总线和/或网络而彼此连接。

[0197] 上述关于特征重要性确定方法所涉及的操作可被描述为各种互联或耦合的功能块或功能示图。然而，这些功能块或功能示图可被均等地集成为单个的逻辑装置或按照非确切的边界进行操作。

[0198] 具体说来，如上所述，根据本发明示例性实施例的确定机器学习样本的各个特征的重要性的计算装置可包括存储部件和处理器，存储部件中存储有计算机可执行指令集合，当所述计算机可执行指令集合被所述处理器执行时，执行下述步骤：(A)获取历史数据记录，其中，所述历史数据记录包括关于机器学习问题的标记和用于生成机器学习样本的各个特征的至少一个属性信息；(B)利用获取的历史数据记录，训练至少一个特征池模型，其中，特征池模型是指基于所述各个特征之中的至少一部分特征来提供关于机器学习问题的预测结果的机器学习模型；(C)获取所述至少一个特征池模型的效果，并根据获取的所述至少一个特征池模型的效果来确定所述各个特征的重要性，其中，在步骤(B)中，通过对所

述至少一部分特征之中的至少一个连续特征执行离散化运算来训练特征池模型。

[0199] 应注意，以上已经结合图2到图5描述了根据本发明示例性实施例的特征重要性确定方法的各处理细节，这里将不再赘述计算装置执行各步骤时的处理细节。

[0200] 以上已经描述了本发明的各示例性实施例，应理解，上述描述仅是示例性的，并非穷尽性的，并且本发明也不限于所披露的各示例性实施例。在不偏离本发明的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。因此，本发明的保护范围应该以权利要求的范围为准。

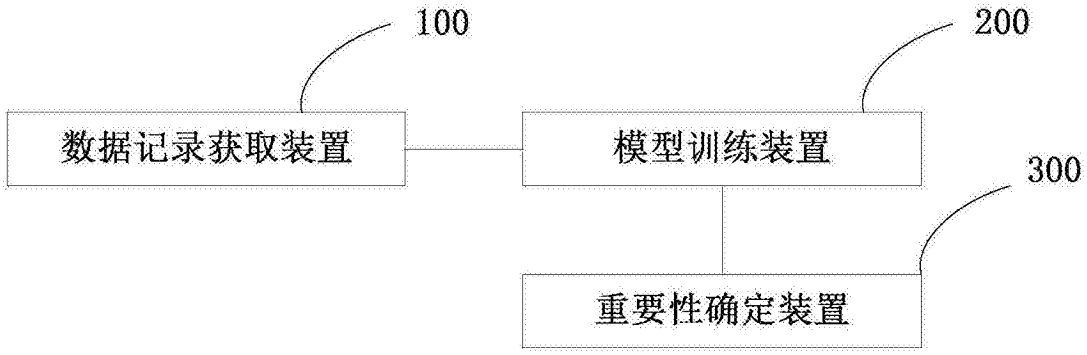

100
200
300
数据记录获取装置
模型训练装置
重要性确定装置

图1

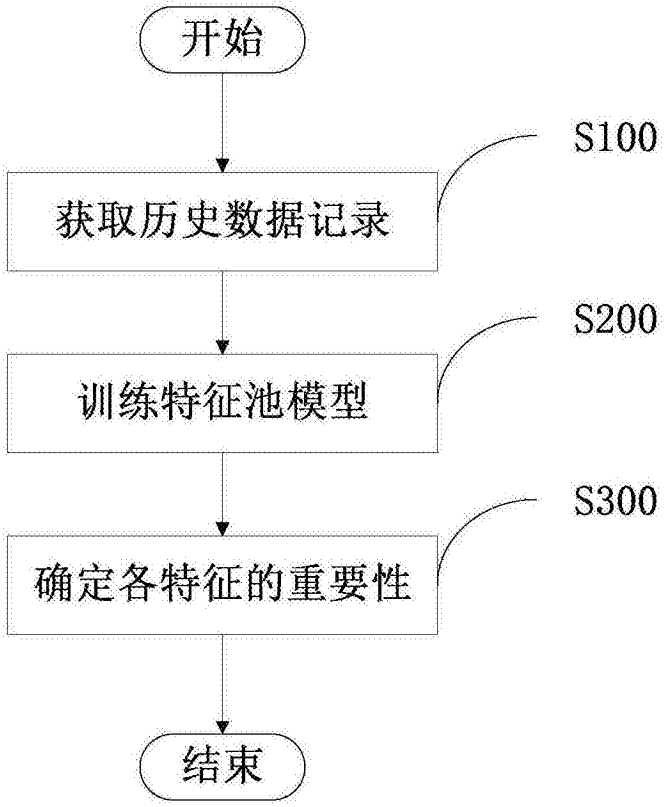

开始
S100
获取历史数据记录
S200
训练特征池模型
S300
确定各特征的重要性
结束

图2

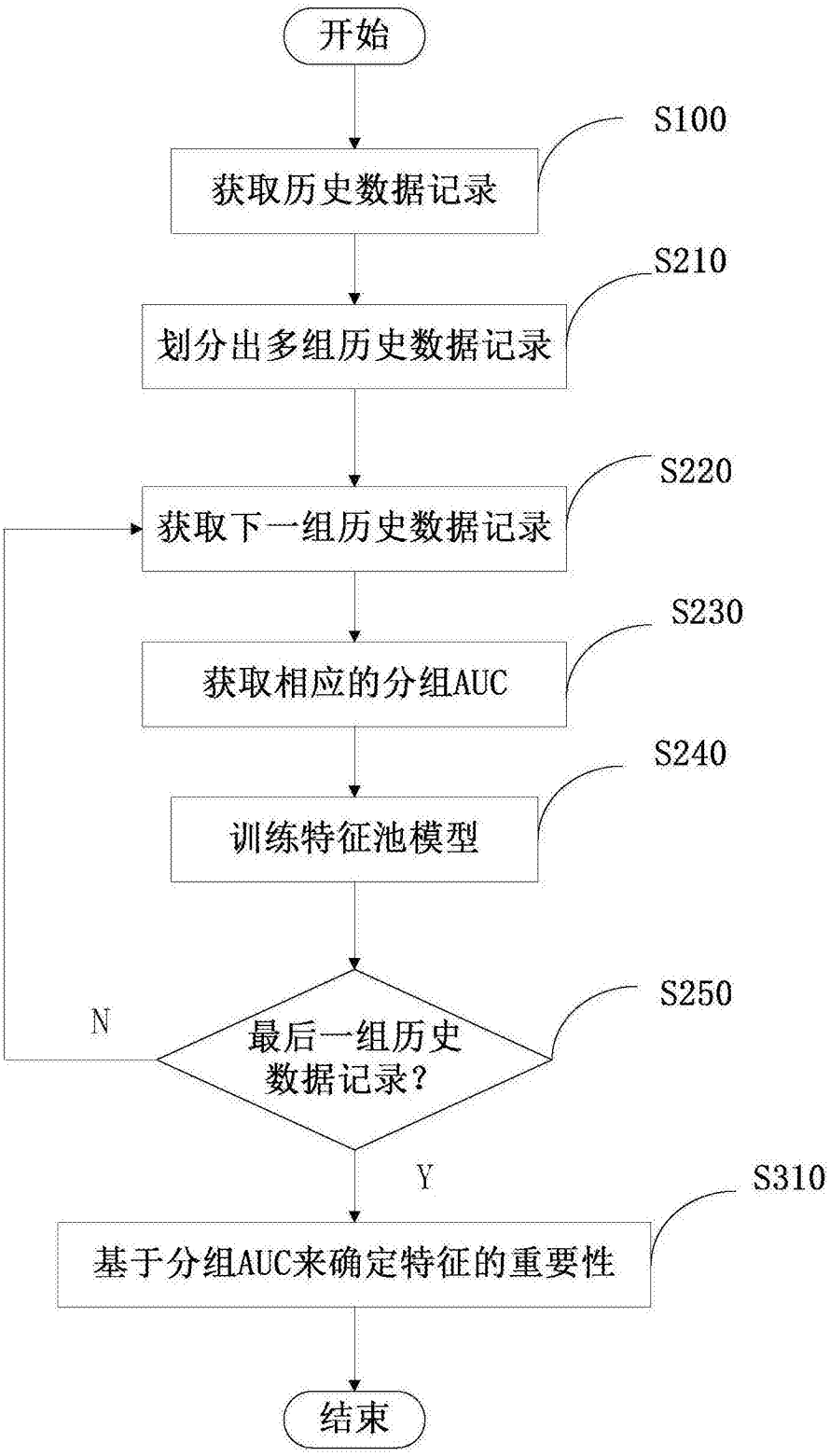

开始
S100
获取历史数据记录
S210
划分出多组历史数据记录
S220
获取下一组历史数据记录
S230
获取相应的分组AUC
S240
训练特征池模型
S250
最后一组历史数据记录？
N
Y
S310
基于分组AUC来确定特征的重要性
结束

图3

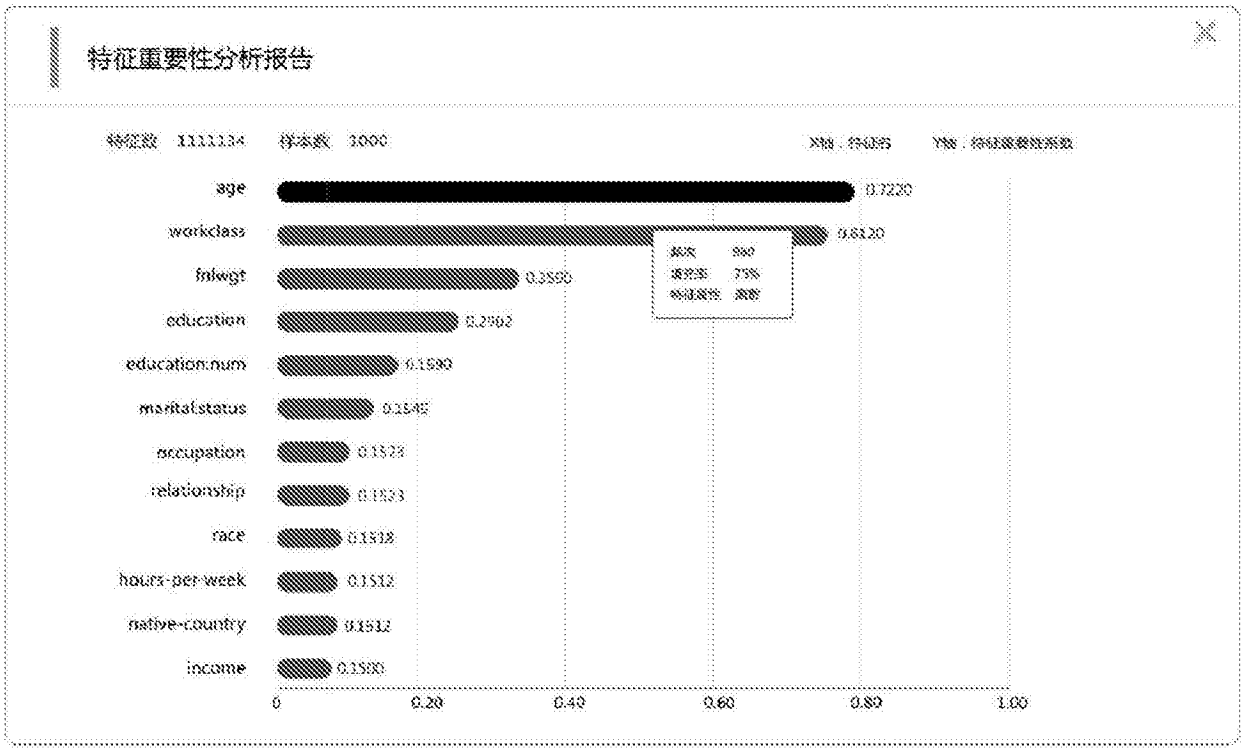

特征重要性分析报告
age
workclass
fnlwgt
education
education-num
marital-status
occupation
relationship
race
hours-per-week
native-country
income
0.7220
0.6120
0.3590
0.2902
0.1590
0.1545
0.1523
0.1523
0.1518
0.1512
0.1512
0.1500
0
0.20
0.40
0.60
0.80
1.00

图4

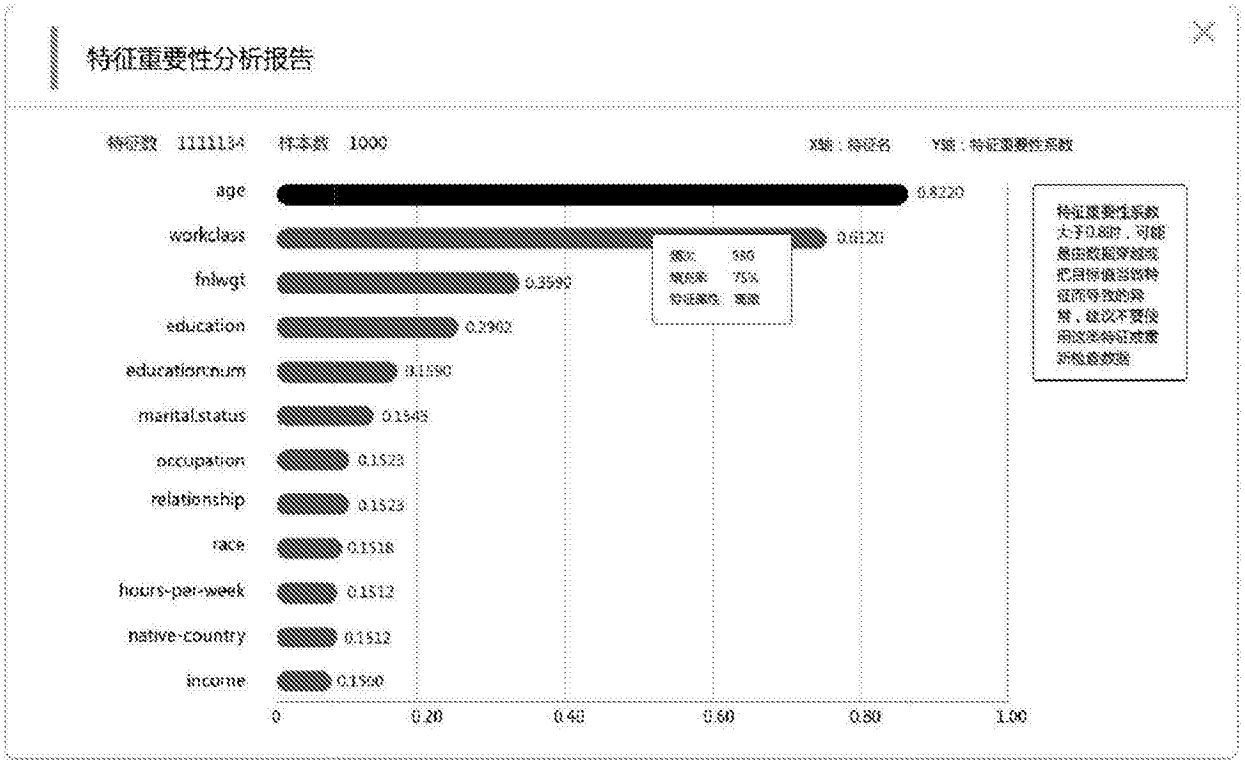

特征重要性分析报告
age
workclass
fnlwgt
education
education-num
marital-status
occupation
relationship
race
hours-per-week
native-country
income
0.8220
0.6120
0.3590
0.2902
0.1590
0.1545
0.1523
0.1523
0.1518
0.1512
0.1512
0.1500
0
0.20
0.40
0.60
0.80
1.00

图5