(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 107316082 A
(43)申请公布日 2017.11.03

(21)申请号 201710469134.1

(22)申请日 2017.06.15

(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 戴文渊 陈雨强 杨强 罗远飞 涂威威

(74)专利代理机构 北京铭硕知识产权代理有限公司 11286
代理人 张云珠 曾世骁

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

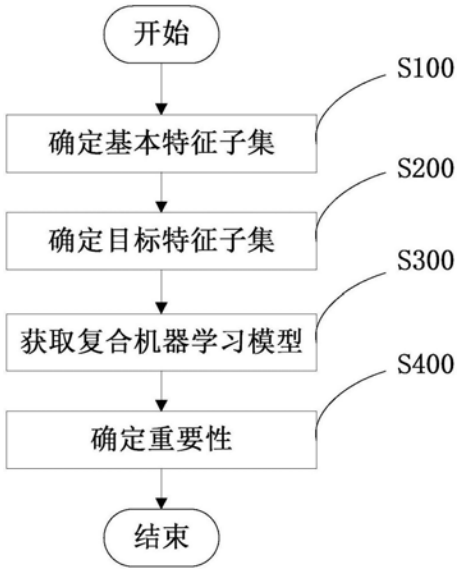
权利要求书1页 说明书10页 附图2页

(54)发明名称

用于确定机器学习样本的特征重要性的方法及系统

(57)摘要

提供了一种用于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)确定机器学习样本的基本特征子集;(B)确定机器学习样本的重要性待确定的多个目标特征子集;(C)针对所述多个目标特征子集之中的每一个目标特征子集,获取相应的复合机器学习模型,其中,所述复合机器学习模型包括根据提升框架训练而成的基本子模型和附加子模型,其中,基本子模型基于基本特征子集训练而成,附加子模型基于所述每一个目标特征子集训练而成;以及(D)根据复合机器学习模型的效果来确定所述多个目标特征子集的重要性。根据所述方法和系统,能够以较低的运算代价有效地得出各个目标特征子集的重要性。



CN 107316082 A

1.一种用于确定机器学习样本的特征重要性的方法,包括:

(A)确定机器学习样本的基本特征子集,其中,基本特征子集包括至少一个基本特征;

(B)确定机器学习样本的重要性待确定的多个目标特征子集,其中,每一个目标特征子集包括至少一个目标特征;

(C)针对所述多个目标特征子集之中的每一个目标特征子集,获取相应的复合机器学习模型,其中,所述复合机器学习模型包括根据提升框架训练而成的基本子模型和附加子模型,其中,基本子模型基于基本特征子集训练而成,附加子模型基于所述每一个目标特征子集训练而成;以及

(D)根据复合机器学习模型的效果来确定所述多个目标特征子集的重要性。

2.如权利要求1所述的方法,其中,在步骤(D)中,根据复合机器学习模型在相同数据集上的效果之间的差异来确定所述多个目标特征子集的重要性。

3.如权利要求1所述的方法,其中,所述目标特征基于基本特征而产生。

4.如权利要求1所述的方法,其中,所述目标特征为通过对至少一个基本特征进行组合而得到的组合特征。

5.如权利要求1所述的方法,其中,在步骤(C)中,通过并行地训练多个复合机器学习模型来获取与每一个目标特征子集相应的复合机器学习模型。

6.如权利要求1所述的方法,其中,目标特征子集包括通过对至少一个基本特征进行组合而得到的一个组合特征,并且,所述方法还包括:(E)以图形化方式向用户展示确定的各个组合特征的重要性。

7.如权利要求1所述的方法,其中,在步骤(C)中,通过在固定已经训练出的基本子模型的情况下训练附加子模型来获取相应的复合机器学习模型。

8.一种用于确定机器学习样本的特征重要性的介质,其中,在所述计算机可读介质上记录有用于执行如权利要求1到7中的任一权利要求所述的方法的计算机程序。

9.一种用于确定机器学习样本的特征重要性的计算装置,包括存储部件和处理器,其中,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,执行如权利要求1到7中的任一权利要求所述的方法。

10.一种用于确定机器学习样本的特征重要性的系统,包括:

基本特征子集确定装置,用于确定机器学习样本的基本特征子集,其中,基本特征子集包括至少一个基本特征;

目标特征子集确定装置,用于确定机器学习样本的重要性待确定的多个目标特征子集,其中,每一个目标特征子集包括至少一个目标特征;

复合机器学习模型获取装置,用于针对所述多个目标特征子集之中的每一个目标特征子集,获取相应的复合机器学习模型,其中,所述复合机器学习模型包括根据提升框架训练而成的基本子模型和附加子模型,其中,基本子模型基于基本特征子集训练而成,附加子模型基于所述每一个目标特征子集训练而成;以及

重要性确定装置,用于根据复合机器学习模型的效果来确定所述多个目标特征子集的重要性。

用于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] 本发明的示例性实施例总体说来涉及人工智能领域，更具体地说，涉及一种用于确定机器学习样本的特征重要性的方法及系统。

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，而为了从海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”。通过对数据记录的属性信息进行诸如特征工程等处理，可产生包括各种特征的机器学习样本。

[0004] 实践中，机器学习模型的预测效果与模型的选择、可用的数据和样本特征的提取均有关系。此外，应用机器学习技术时还需要面对计算资源有限、样本数据不足等客观问题。因此，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特征或特征组合的重要程度。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，这些因素都导致特征提取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衡量机器学习样本特征重要性的缺陷。

[0007] 根据本发明的示例性实施例，提供一种用于确定机器学习样本的特征重要性的方法，包括：(A)确定机器学习样本的基本特征子集，其中，基本特征子集包括至少一个基本特征；(B)确定机器学习样本的重要性待确定的多个目标特征子集，其中，每一个目标特征子集包括至少一个目标特征；(C)针对所述多个目标特征子集之中的每一个目标特征子集，获取相应的复合机器学习模型，其中，所述复合机器学习模型包括根据提升框架训练而成的基本子模型和附加子模型，其中，基本子模型基于基本特征子集训练而成，附加子模型基于所述每一个目标特征子集训练而成；以及(D)根据复合机器学习模型的效果来确定所述多个目标特征子集的重要性。

[0008] 可选地，在所述方法中，在步骤(D)中，根据复合机器学习模型在相同数据集上的效果之间的差异来确定所述多个目标特征子集的重要性。

[0009] 可选地，在所述方法中，复合机器学习模型的效果包括复合机器学习模型的AUC。

[0010] 可选地，在所述方法中，所述目标特征基于基本特征而产生。

[0011] 可选地，在所述方法中，所述目标特征为通过对至少一个基本特征进行组合而得到的组合特征。

[0012] 可选地，在所述方法中，在步骤(C)中，通过并行地训练多个复合机器学习模型来获取与每一个目标特征子集相应的复合机器学习模型。

[0013] 可选地，在所述方法中，目标特征子集包括通过对至少一个基本特征进行组合而得到的一个组合特征，并且，所述方法还包括：(E)以图形化方式向用户展示确定的各个组合特征的重要性。

[0014] 可选地，在所述方法中，在步骤(C)中，通过在固定已经训练出的基本子模型的情况下训练附加子模型来获取相应的复合机器学习模型。

[0015] 可选地，在所述方法中，基本子模型和附加子模型的类型相同。

[0016] 根据本发明的另一示例性实施例，提供一种用于确定机器学习样本的特征重要性的介质，其中，在所述计算机可读介质上记录有用于执行上述方法的计算机程序。

[0017] 根据本发明的另一示例性实施，提供一种用于确定机器学习样本的特征重要性的计算装置，包括存储部件和处理器，其中，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，执行上述方法。

[0018] 根据本发明的另一示例性实施例，提供一种用于确定机器学习样本的特征重要性的系统，包括：基本特征子集确定装置，用于确定机器学习样本的基本特征子集，其中，基本特征子集包括至少一个基本特征；目标特征子集确定装置，用于确定机器学习样本的重要性待确定的多个目标特征子集，其中，每一个目标特征子集包括至少一个目标特征；复合机器学习模型获取装置，用于针对所述多个目标特征子集之中的每一个目标特征子集，获取相应的复合机器学习模型，其中，所述复合机器学习模型包括根据提升框架训练而成的基本子模型和附加子模型，其中，基本子模型基于基本特征子集训练而成，附加子模型基于所述每一个目标特征子集训练而成；以及重要性确定装置，用于根据复合机器学习模型的效果来确定所述多个目标特征子集的重要性。

[0019] 可选地，在所述系统中，重要性确定装置根据复合机器学习模型在相同数据集上的效果之间的差异来确定所述多个目标特征子集的重要性。

[0020] 可选地，在所述系统中，复合机器学习模型的效果包括复合机器学习模型的AUC。

[0021] 可选地，在所述系统中，所述目标特征基于基本特征而产生。

[0022] 可选地，在所述系统中，所述目标特征为通过对至少一个基本特征进行组合而得到的组合特征。

[0023] 可选地，在所述系统中，复合机器学习模型获取装置通过并行地训练多个复合机器学习模型来获取与每一个目标特征子集相应的复合机器学习模型。

[0024] 可选地，在所述系统中，目标特征子集包括通过对至少一个基本特征进行组合而得到的一个组合特征，并且，所述系统还包括：显示装置，以图形化方式向用户展示确定的各个组合特征的重要性。

[0025] 可选地，在所述系统中，复合机器学习模型获取装置通过在固定已经训练出的基本子模型的情况下训练附加子模型来获取相应的复合机器学习模型。

[0026] 可选地，在所述系统中，基本子模型和附加子模型的类型相同。

[0027] 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，每个复合机器学习模型被构建为包括基于提升框架的基本子模型(与基本特征子集对应)和附加子模型(与重要性待确定的某个目标特征子集对应)，相应地，根据各个复合机器学习模型的效果，能够以较低的运算代价有效地得出各个目标特征子集的重要性。

附图说明

[0028] 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：

[0029] 图1示出根据本发明示例性实施例的用于确定机器学习样本的特征重要性的系统的框图；

[0030] 图2示出根据本发明示例性实施例的用于确定机器学习样本的特征重要性的方法的流程图；

[0031] 图3示出根据本发明示例性实施例的训练复合机器学习模型的方法的流程图；以及

[0032] 图4示出根据本发明示例性实施例的测试复合机器学习模型的方法的流程图。

具体实施方式

[0033] 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。

[0034] 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。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，本发明的示例性实施例在训练和应用机器学习模型的过程中，还可利用统计算法、业务规则和/或专家知识等，以进一步提高机器学习的效果。

[0035] 本发明的示例性实施例涉及如何衡量机器学习样本的特征重要性，在本发明的示例性实施例中，复合机器学习模型被构建为包括基于提升框架的基本子模型和附加子模型，其中，所有复合机器学习模型的基本子模型均对应于同样的基本特征子集，而每个复合机器学习模型的附加子模型对应于各自不同的目标特征子集，因此，可通过比较各个复合机器学习模型的效果来衡量相应目标特征子集的重要性。这里，由于目标特征子集可包括一个或多个目标特征，因此，本发明的示例性实施例既可以衡量多个目标特征之间的重要性，也可以衡量多组目标特征之间的重要性。

[0036] 图1示出根据本发明示例性实施例的用于确定机器学习样本的特征重要性的系统。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相关特征的重要性，这些重要性信息有助于进行模型训练和/或模型解释。

[0037] 参照图1，基本特征子集确定装置100用于确定机器学习样本的基本特征子集，其中，基本特征子集包括至少一个基本特征。根据本发明的示例性实施例，基本特征子集将固定地应用于所有复合机器学习模型中的基本子模型，这里，基本特征子集确定装置100可将任何基于数据记录的属性信息产生的特征作为基本特征。例如，基本特征子集确定装置100可将数据记录的至少一部分属性信息直接作为基本特征。此外，作为示例，基本特征子集确定装置100可考虑实际的机器学习问题，基于测试计算或根据业务人员指定来确定相对重要或基本的特征作为基本特征。

[0038] 目标特征子集确定装置200用于确定机器学习样本的重要性待确定的多个目标特征子集，其中，每一个目标特征子集包括至少一个目标特征。这里，每个目标特征子集可包括一个或多个目标特征，并且，目标特征子集确定装置200可将任何基于数据记录的属性信息产生的特征作为目标特征。

[0039] 具体说来，当期望确定多个特征之间的重要性时，目标特征子集确定装置200可将所述多个特征之中的每一个特征作为目标特征子集。当期望确定多组特征之间的重要性时，目标特征子集确定装置200可将所述多组特征之中的每一组特征作为目标特征子集。根据本发明的示例性实施例，每个复合机器学习模型的附加子模型对应于各自的目标特征子集，相应地，可根据多个复合机器学习模型的效果差异来衡量不同目标特征子集在预测时所起到的作用大小。

[0040] 在获知了目标特征子集的重要性之后，可在模型的训练和/或解释等方面利用这样的重要性信息。作为示例，在所有目标特征子集均只包括单个目标特征的情况下，可通过衡量所有目标特征的重要性来筛选出较为重要的一个或多个特征，以作为训练样本的特征。作为另一示例，在目标特征子集分别包括不同目标特征的组合时，可衡量出这些特征组合的不同表现，以选取最优组合作为训练样本的特征。应注意，本发明的示例性实施例并不限制应用重要性确定结果的具体方式。

[0041] 可以看出，根据本发明示例性实施例的基本特征和目标特征是一个相对的概念，使得在保持基本特征不变的情况下，衡量各个复合机器学习模型由于引入不同的目标特征子集所带来的效果差异。基于上述构思，可根据具体情况，采用任何适当的方式来设计基本特征子集和各个目标特征子集。

[0042] 复合机器学习模型获取装置300用于针对所述多个目标特征子集之中的每一个目标特征子集，获取相应的复合机器学习模型，其中，所述复合机器学习模型包括根据提升框架训练而成的基本子模型和附加子模型，其中，基本子模型基于基本特征子集训练而成，附加子模型基于所述每一个目标特征子集训练而成。

[0043] 根据本发明的示例性实施例，对于每一个目标特征子集，需获取对应的复合机器学习模型。这里，复合机器学习模型获取装置300可自身完成复合机器学习模型的训练，也可从外部获取已经训练好的复合机器学习模型。这里，复合机器学习模型包括根据提升框架(例如，梯度提升框架)训练而成的基本子模型和附加子模型，其中，基本子模型和附加子模型可以是类型相同的模型，例如，基本子模型和附加子模型可以都是线性模型(例如，对数几率回归模型)，此外，基本子模型和附加子模型也可以具有不同的类型。这里，各个复合机器学习模型的提升框架可以是相同的，即，各个复合机器学习模型具有相同类型的基本子模型和相同类型的附加子模型，区别仅在于附加子模型所依据的目标特征子集不一样。

[0044] 重要性确定装置400用于根据复合机器学习模型的效果来确定所述多个目标特征子集的重要性。如上所述，复合机器学习模型的效果可用于衡量其所对应的目标特征子集的重要性。这里，重要性确定装置400可通过测试不同复合机器学习模型在相同数据集上的表现来反映各个目标特征子集的重要性。

[0045] 作为示例，图1所示的系统还可包括显示装置(未示出)，用于以图形化的形式向用户展示特征重要性的确定结果。例如，可将目标特征子集的重要性展示为图形或表格，以便用户更好地进行特征工程或更直观地理解模型。此外，图1所示的系统还可包括输入装置(未示出)，用于感测用户为了指定特征处理方式等而进行的输入操作。

[0046] 以下将参照图2来描述根据本发明示例性实施例的用于确定机器学习样本的特征重要性的方法。这里，作为示例，图2所示的方法可由图1所示的系统来执行，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。

[0047] 为了描述方便，假设图2所示的方法由图1所示的系统来执行。参照图2，在步骤S100中，由基本特征子集确定装置100确定机器学习样本的基本特征子集，其中，基本特征子集包括至少一个基本特征。作为示例，基本特征子集确定装置100可根据预设的特征提取方式来确定基本特征子集所包括的各个基本特征，例如，可将数据记录的预定属性信息按照设定的方式处理为相应的基本特征。这里，可借助属性测试手段或根据业务经验来确定所述预定属性信息和/或相应的处理方式。此外，基本特征子集确定装置100可根据用户的交互操作来确定基本特征子集所包括的各个基本特征，例如，在诸如机器学习平台的软件系统中，用户可通过相应的交互操作来手动选取基本特征。优选地，上述两种方式还可进行结合，例如，用户可通过诸如软件系统的操作界面设置基本特征的处理方式，包括作为特征来源的属性信息、提取方式(例如，直接提取、组合提取、运算提取等)、相关参数等。

[0048] 在步骤S200中，由目标特征子集确定装置200确定机器学习样本的重要性待确定的多个目标特征子集，其中，每一个目标特征子集包括至少一个目标特征。作为示例，目标特征子集确定装置200可根据预设的特征提取方式来确定各个目标特征子集所包括的目标特征，例如，可将数据记录的预定属性信息按照设定的方式处理为相应的目标特征。这里，可借助属性测试手段或根据业务经验来确定所述预定属性信息和/或相应的处理方式。此外，目标特征子集确定装置200可根据用户的交互操作来确定目标特征子集所包括的各个目标特征，例如，在诸如机器学习平台的软件系统中，用户可通过相应的交互操作来手动选取各个目标特征子集所包括的目标特征。优选地，上述两种方式还可进行结合，例如，用户可通过诸如软件系统的操作界面设置目标特征的处理方式，包括作为特征来源的属性信息、提取方式(例如，直接提取、组合提取、运算提取等)、相关参数等。

[0049] 根据本发明的示例性实施例，所述目标特征可基于基本特征而产生。例如，可通过对至少一个基本特征进行某种变换而得到相应的目标特征。作为示例，所述目标特征可以为通过对至少一个基本特征进行组合而得到的组合特征。在组合基本特征时，还可先对相关基本特征进行附加的变换(例如，指数运算、离散化等)。例如，在基本特征子集包括特征a、特征b、特征c、特征d等的情况下，组合特征可以是上述特征的直接组合，例如，组合特征可以是上述一部分特征的笛卡尔积，或者，特别地，组合特征可以是单个的上述特征本身；此外，组合特征也可以是上述特征的算术运算结果(例如，a^2、d^2等)的组合(例如，可以是上述算术运算结果的笛卡尔积或其本身)。通过这种方式来划分基本特征和目标特征并结

合复合机器学习模型的框架，可有效地衡量机器学习样本的一些较为复杂的目标特征(例如，组合特征)是否适合作为最终使用的特征。

[0050] 在步骤S300中，由复合机器学习模型获取装置300针对所述多个目标特征子集之中的每一个目标特征子集，获取相应的复合机器学习模型，其中，所述复合机器学习模型包括根据提升框架训练而成的基本子模型和附加子模型，其中，基本子模型基于基本特征子集训练而成，附加子模型基于所述每一个目标特征子集训练而成。

[0051] 作为示例，复合机器学习模型获取装置300可从外部获取已经训练好的复合机器学习模型，为此，复合机器学习模型获取装置300需要首先将之前确定的基本特征子集和各个目标特征子集的具体提取方式通知外部的模型训练装置(未示出，可位于图1所示的系统之内或图1所示的系统之外)，以便所述外部的模型训练装置可按照相应的特征设计来构建训练样本，进而训练出复合机器学习模型。

[0052] 作为另一示例，复合机器学习模型获取装置300本身可执行复合机器学习模型的训练过程。以下结合图3来描述根据本发明示例性实施例的训练复合机器学习模型的方法。

[0053] 参照图3，在步骤S310中，可获取训练数据记录。这些训练数据记录可由任何方以任何方式来产生，例如，可以是在线生成或收集的数据、预先生成或存储的数据、也可以是从外部接收的数据。这些数据的属性信息可涉及客户信息，例如，身份、学历、职业、资产、联系方式等信息。或者，这些数据的属性信息也可涉及业务相关项目的信息，例如，关于买卖合同的交易额、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数据的属性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。实际上，任何能够通过对其进行机器学习的信息数据均可应用于本发明的示例性实施例。

[0054] 这里，可获取不同来源(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)的结构化或非结构化数据，例如，文本数据或数值数据等。这些数据可从外部输入到复合机器学习模型获取装置300，或者由复合机器学习模型获取装置300根据已有的数据来自动生成，或者可由复合机器学习模型获取装置300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复合机器学习模型获取装置300从外部数据源获取相应的数据。这里，获取的数据可被复合机器学习模型获取装置300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。应注意，复合机器学习模型获取装置300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。

[0055] 接下来，在步骤S320中，可基于训练数据记录的属性信息，按照之前确定的基本特征子集和/或目标特征子集来生成复合机器学习模型的训练样本。如上所述，根据本发明示例性实施例的复合机器学习模型包括根据提升框架的基本子模型和附加子模型。相应地，本领域技术人员应理解，作为示例，对于每一个复合机器学习模型而言，可首先训练出基本子模型，然后训练出附加子模型，相应地，可构建用于训练出基本子模型的训练样本(其包括基本特征子集和标记(label)部分)和用于随后训练出附加子模型的训练样本(其包括基本特征子集、目标特征子集和标记部分)。

[0056] 作为示例，可根据基本特征子集和目标特征子集的具体设置，通过对训练数据记

录的属性信息进行筛选、分组或进一步附加处理等而得到相应特征。根据本发明的示例性实施例，可按照任何适当的方式来生成相应特征，例如，可考虑属性信息的内容、含义、取值连续性、取值范围、取值空间规模、缺失性、重要性等因素，或者，可结合复合机器学习模型中的子模型特点等。

[0057] 根据本发明的示例性实施例，可基于基本特征子集中的基本特征来产生目标特征子集中的目标特征，也就是说，目标特征基于基本特征而产生。例如，可将基本特征的组合作为目标特征。这里，可通过对基本特征进行任何适当的变换来得到目标特征。随着目标特征经由附加子模型而引入到机器学习模型中，能够相应地影响机器学习模型的效果。

[0058] 在步骤S330中，可利用生成的训练样本来训练复合机器学习模型。根据本发明的示例性实施例，在每个复合机器学习模型中，基本子模型与附加子模型之间基于提升框架训练而成。

[0059] 具体说来，可根据提升框架(例如，梯度提升框架)来训练复合机器学习模型所包括的基本子模型和附加子模型，这两个子模型可具有相同或不同的模型类型。这里，针对每一个复合机器学习模型，可基于载入的模型训练配置来分阶段地训练出基本子模型和附加子模型，具体说来，在第一阶段训练基本子模型时，可根据配置的参数来执行初始化处理，并利用由基本特征子集与标记部分组成的训练样本来训练基本子模型。在此基础上，提升框架下的复合机器学习模型可表示为基本子模型和附加子模型的拼接结果，该结果可对应于一个相对较强的模型。相应地，在训练出基本子模型之后，可利用由基本特征子集、目标特征子集连同标记部分组成的训练样本来训练附加子模型。

[0060] 假设单个复合机器学习模型表示为F，这里，F可由基本子模型f_{base}和附加子模型f_{add}组成，假设输入的训练数据记录表示为x，在按照确定的基本特征子集和目标特征子集经过相应的特征处理之后，基本子模型f_{base}对应的样本部分的特征为x^b，附加子模型f_{add}对应的样本部分的特征为x^a。相应地，可按照以下的等式来构建复合机器学习模型F：

[0061] $F(x)=f_{base}(x^b)+f_{add}(x^a)$。

[0062] 然而，应注意，基本子模型和附加子模型除了可基于相同的训练数据记录集训练而成之外，还可基于不同的训练数据记录集训练而成。例如，上述两种子模型均可基于全体训练数据记录训练而成，或者，也可分别基于从全体训练数据记录中采样的一部分训练数据记录训练而成。作为示例，可根据预设的采样策略为基本子模型和附加子模型分配相应的训练数据记录，例如，可将较多的训练数据记录分配给基本子模型，而将较少的训练数据记录分配给附加子模型，这里，不同子模型分配的训练数据记录之间可具有一定比例的交集或者完全没有交集。通过根据采样策略来确定各个子模型所使用的训练数据记录，可进一步提升整个机器学习模型的效果。

[0063] 根据本发明的示例性实施例，可通过并行地训练多个复合机器学习模型来获取与每一个目标特征子集相应的复合机器学习模型。作为示例，在训练附加子模型时，基本子模型的系数可固定不变。也就是说，通过在固定已经训练出的基本子模型的情况下训练附加子模型来获取相应的复合机器学习模型。在这种情况下，可大大降低并行训练时的运算量，降低了内存需求。

[0064] 以上列出了子模型的示例性训练方式，然而，应理解，本发明的示例性实施例并不受限于上述示例。

[0065] 再次参照回图2,在获取了复合机器学习模型之后,在步骤S400中,由重要性确定装置400根据复合机器学习模型的效果来确定所述多个目标特征子集的重要性。这里,作为示例,重要性确定装置400可通过执行相应处理来亲自确定各个复合机器学习模型的效果,也可从与其连接的其他方接收各个复合机器学习模型的效果。例如,为了确定复合机器学习模型的效果,可在逐步地训练复合机器学习模型的同时获取模型效果。具体说来,可将训练样本划分为多组以逐步地训练复合机器学习模型,并且,可在训练过程中,使用当前训练出的复合机器学习模型来针对下一组训练样本执行预测以得到与所述下一组训练样本相应的分组效果,并综合各个分组效果来得到复合机器学习模型的总效果,其中,在得到所述下一组训练样本的分组效果之后,利用所述下一组训练样本来继续训练当前模型。又例如,为了确定复合机器学习模型的效果,可在多个复合机器学习模型的训练完成之后,通过将这些复合机器学习模型应用于相应的测试数据集来获取各个复合机器学习模型的效果,其中,所述测试数据集既可以包括用于模型训练的训练数据记录,也可以包括除了训练数据记录之外的其他历史数据记录。这里,可根据复合机器学习模型在相同数据集上的效果之间的差异来确定多个目标特征子集的重要性;或者,也可根据复合机器学习模型在不同数据集上的效果之间的差异来确定多个目标特征子集的重要性。

[0066] 具体说来,复合机器学习模型在测试集上的表现可作为该复合机器学习模型的预测效果,而这一预测效果可用于衡量所述复合机器学习模型的目标特征子集的预测能力。通过衡量不同复合机器学习模型在原始测试数据集上的效果差异,可综合得出机器学习样本的各个目标特征子集的重要性。

[0067] 这里,作为示例,复合机器学习模型的效果可包括复合机器学习模型的AUC(ROC(受试者工作特征,Receiver Operating Characteristic)曲线下的面积,Area Under ROC Curve)或对率损失(logistic loss)。

[0068] 图4示出根据本发明示例性实施例的测试复合机器学习模型的方法的流程图。通过执行图4所示的方法,可获得各个复合机器学习模型的效果。

[0069] 具体说来,在步骤S410中,可获取测试数据记录。这里,作为示例,测试数据记录可以是除了训练数据记录以外的其他历史数据记录。

[0070] 接着,在步骤S420中,可针对每个复合机器学习模型,根据相应的基本特征子集和目标特征子集的具体设置,对测试数据记录进行特征工程处理,以得到所述每个复合机器学习模型的测试样本。这里,每个复合机器学习模型的测试样本可基于同样的测试数据集。

[0071] 在步骤S430中,获取每个复合机器学习模型针对相应的测试样本所产生的预测结果,从而基于预测结果来得到各个复合机器学习模型的效果。

[0072] 应理解,图4所示的示例仅用于说明本发明的示例性实施,而不是为了进行限制,例如,本发明的示例性实施例还可在训练复合机器学习模型的过程中,利用后续训练数据记录来逐步测试当前训练出的复合机器学习模型的效果,并在模型训练完成之后综合得出复合机器学习模型的整体效果。

[0073] 在确定了各个复合机器学习模型的效果之后,可根据效果之间的差异来确定相应目标特征子集的重要性。对于如何应用目标特征子集的重要性确定结果,本发明的示例性实施例并不受限。例如,可将每个目标特征子集均设置为仅包括通过对至少一个基本特征进行组合而得到的一个组合特征,相应地,在确定了各个组合特征的重要性之后,还可通过

图形化方式向用户展示确定的各个组合特征的重要性。

[0074] 应理解，图1所示出的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。

[0075] 以上参照图1和图2描述了根据本发明示例性实施例的用于确定机器学习样本的特征重要性的系统和方法。应理解，上述方法可通过记录在计算可读介质上的程序来实现，相应地，根据本发明的示例性实施例，可提供一种用于确定机器学习样本的特征重要性的介质，其中，在所述计算机可读介质上记录有用于执行以下方法步骤的计算机程序：(A)确定机器学习样本的基本特征子集，其中，基本特征子集包括至少一个基本特征；(B)确定机器学习样本的重要性待确定的多个目标特征子集，其中，每一个目标特征子集包括至少一个目标特征；(C)针对所述多个目标特征子集之中的每一个目标特征子集，获取相应的复合机器学习模型，其中，所述复合机器学习模型包括根据提升框架训练而成的基本子模型和附加子模型，其中，基本子模型基于基本特征子集训练而成，附加子模型基于所述每一个目标特征子集训练而成；以及(D)根据复合机器学习模型的效果来确定所述多个目标特征子集的重要性。

[0076] 上述计算机可读介质中的计算机程序可在诸如客户端、主机、代理装置、服务器等计算机设备中部署的环境中运行，应注意，所述计算机程序还可用于执行除了上述步骤以外的附加步骤或者在执行上述步骤时执行更为具体的处理，这些附加步骤和进一步处理的内容已经参照图1到图4进行了描述，这里为了避免重复将不再进行赘述。

[0077] 应注意，根据本发明示例性实施例的特征重要性确定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与各步骤相应，使得整个系统通过专门的软件包(例如，lib库)而被调用，以实现相应的预测功能。

[0078] 另一方面，图1所示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。

[0079] 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，执行用于确定机器学习样本的特征重要性的方法。

[0080] 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。

[0081] 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。

[0082] 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。

[0083] 根据本发明示例性实施例的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。

[0084] 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。

[0085] 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。

[0086] 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。

[0087] 根据本发明示例性实施例的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。

[0088] 具体说来，如上所述，根据本发明示例性实施例的用于确定机器学习样本的特征重要性的计算装置可包括存储部件和处理器，其中，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)确定机器学习样本的基本特征子集，其中，基本特征子集包括至少一个基本特征；(B)确定机器学习样本的重要性待确定的多个目标特征子集，其中，每一个目标特征子集包括至少一个目标特征；(C)针对所述多个目标特征子集之中的每一个目标特征子集，获取相应的复合机器学习模型，其中，所述复合机器学习模型包括根据提升框架训练而成的基本子模型和附加子模型，其中，基本子模型基于基本特征子集训练而成，附加子模型基于所述每一个目标特征子集训练而成；以及(D)根据复合机器学习模型的效果来确定所述多个目标特征子集的重要性。

[0089] 应注意，以上已经结合图1到图4描述了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。

[0090] 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。

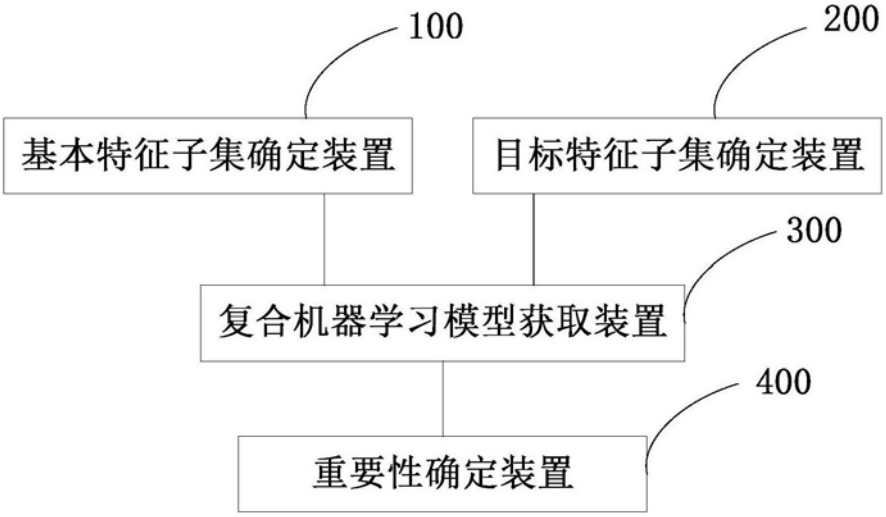

100
200
基本特征子集确定装置
目标特征子集确定装置
300
复合机器学习模型获取装置
400
重要性确定装置

图1

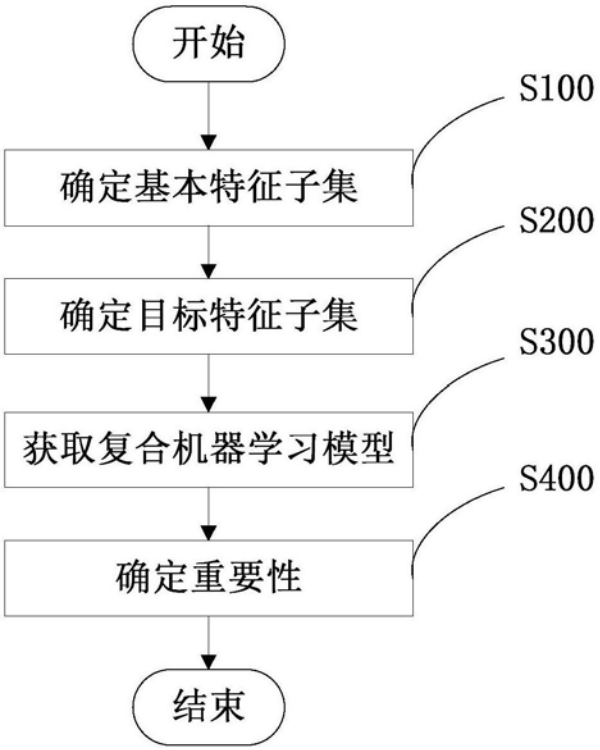

开始
S100
确定基本特征子集
S200
确定目标特征子集
S300
获取复合机器学习模型
S400
确定重要性
结束

图2

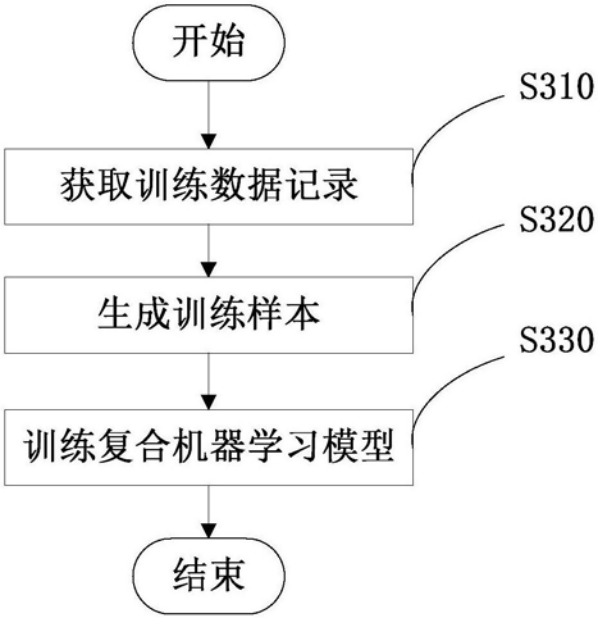

开始
S310
获取训练数据记录
S320
生成训练样本
S330
训练复合机器学习模型
结束

图3

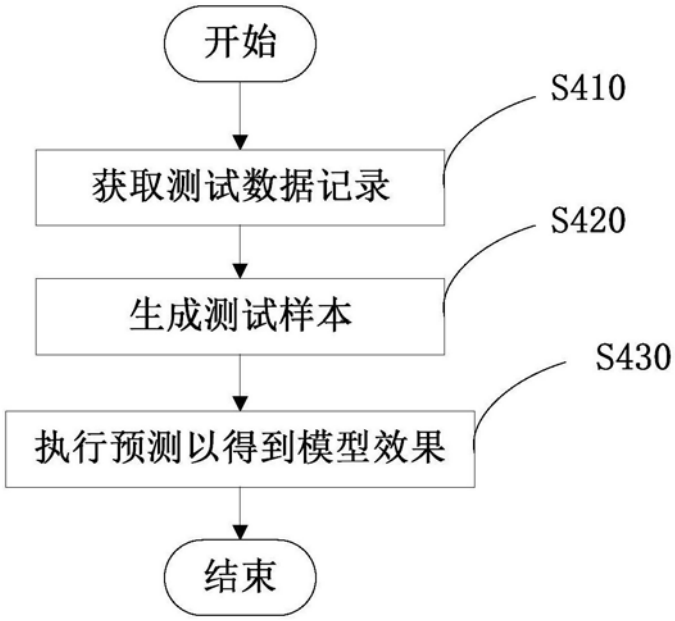

开始
S410
获取测试数据记录
S420
生成测试样本
S430
执行预测以得到模型效果
结束

图4