(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 111797927 A
(43)申请公布日 2020.10.20

(21)申请号 202010641229.9

(22)申请日 2017.09.08

(62)分案原申请数据
201710804531.X 2017.09.08

(71)申请人 第四范式(北京)技术有限公司
地址 100085 北京市海淀区清河中街66号院1号楼九层L0901-1号

(72)发明人 杨强 戴文渊 陈雨强 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 田方 曾世骁

(51)Int.Cl.
G06K 9/62(2006.01)
G06N 20/00(2019.01)
G06F 16/215(2019.01)
G06F 16/31(2019.01)

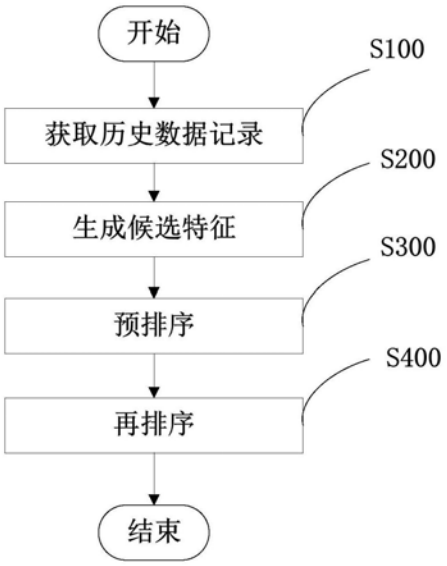
权利要求书2页 说明书17页 附图3页

(54)发明名称

用于确定机器学习样本的重要特征的方法及系统

(57)摘要

提供了一种用于确定机器学习样本的重要特征的方法及系统,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多个属性信息;(B)基于所述多个属性信息生成至少一个候选特征;(C)对所述至少一个候选特征进行重要性的预排序,并根据预排序结果从所述至少一个候选特征中筛选出一部分候选特征以组成候选特征池;以及(D)对候选特征池中的各个候选特征进行重要性的再排序,并根据再排序结果从候选特征池中选择重要性较高的至少一个候选特征作为重要特征。通过特定方式的预排序和再排序从候选特征中筛选出相对重要的特征,从而可在使用较少运算资源的情况下有效地确定重要特征,有助于提升机器学习模型的效果。



CN 111797927 A

1.一种用于确定机器学习样本的重要特征的方法,包括:

(A)获取历史数据记录,其中,所述历史数据记录包括多个属性信息;

(B)基于所述多个属性信息生成至少一个候选特征;

(C)对所述至少一个候选特征进行重要性的预排序,并根据预排序结果从所述至少一个候选特征中筛选出一部分候选特征以组成候选特征池;以及

(D)对候选特征池中的各个候选特征进行重要性的再排序,并根据再排序结果从候选特征池中选择重要性较高的至少一个候选特征作为重要特征。

2.如权利要求1所述的方法,其中,在步骤(C)中,根据预排序结果从所述至少一个候选特征中筛选出重要性较高的候选特征以组成候选特征池。

3.如权利要求1所述的方法,其中,在步骤(C)中,通过以下处理来进行预排序:针对每一个候选特征,得到预排序单特征机器学习模型,基于各个预排序单特征机器学习模型的效果来确定各个候选特征的重要性,其中,预排序单特征机器学习模型对应所述每一个候选特征。

4.如权利要求1所述的方法,其中,在步骤(C)中,通过以下处理来进行预排序:针对每一个候选特征,得到预排序整体机器学习模型,基于各个预排序整体机器学习模型的效果来确定各个候选特征的重要性,其中,预排序整体机器学习模型对应预排序基本特征子集和所述每一个候选特征。

5.如权利要求1所述的方法,其中,在步骤(C)中,通过以下处理来进行预排序:针对每一个候选特征,得到预排序复合机器学习模型,基于各个预排序复合机器学习模型的效果来确定各个候选特征的重要性,其中,预排序复合机器学习模型包括基于提升框架的预排序基本子模型和预排序附加子模型,其中,预排序基本子模型对应预排序基本特征子集,预排序附加子模型对应所述每一个候选特征。

6.如权利要求1所述的方法,其中,在步骤(D)中,通过以下处理来进行再排序:针对候选特征池中的每一个候选特征,得到再排序单特征机器学习模型,基于各个再排序单特征机器学习模型的效果来确定各个候选特征的重要性,其中,再排序单特征机器学习模型对应所述每一个候选特征。

7.如权利要求1所述的方法,其中,在步骤(D)中,通过以下处理来进行再排序:针对候选特征池中的每一个候选特征,得到再排序整体机器学习模型,基于各个再排序整体机器学习模型的效果来确定各个候选特征的重要性,其中,再排序复合机器学习模型对应再排序基本特征子集和所述每一个候选特征。

8.如权利要求1所述的方法,其中,在步骤(D)中,通过以下处理来进行再排序:针对候选特征池中的每一个候选特征,得到再排序复合机器学习模型,基于各个再排序复合机器学习模型的效果来确定各个候选特征的重要性,其中,再排序复合机器学习模型包括基于提升框架的再排序基本子模型和再排序附加子模型,其中,再排序基本子模型对应再排序基本特征子集,再排序附加子模型对应所述每一个候选特征。

9.如权利要求1所述的方法,还包括:(E)检验所述重要特征是否适于作为机器学习样本的特征。

10.一种用于确定机器学习样本的重要特征的系统,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多个属性信

息；

候选特征生成装置，用于基于所述多个属性信息生成至少一个候选特征；

预排序装置，用于对所述至少一个候选特征进行重要性的预排序，并根据预排序结果从所述至少一个候选特征中筛选出一部分候选特征以组成候选特征池；以及

再排序装置，用于对候选特征池中的各个候选特征进行重要性的再排序，并根据再排序结果从候选特征池中选择重要性较高的至少一个候选特征作为重要特征。

用于确定机器学习样本的重要特征的方法及系统

[0001] 本申请是申请日为2017年09月08日、申请号为201710804531.X、题为“用于确定机器学习样本的重要特征的方法及系统”的专利申请的分案申请。

技术领域

[0002] 本发明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重要特征的方法及系统。

背景技术

[0003] 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。

[0004] 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。

[0005] 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(包括特征组合)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。

[0006] 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

[0007] 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。

[0008] 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重要特征的方法，包括：(A)获取历史数据记录，其中，所述历史数据记录包括多个属性信息；(B)基于所述多个属性信息生成至少一个候选特征；(C)对所述至少一个候选特征进行重要性的预排序，并根据预排序结果从所述至少一个候选特征中筛选出一部分候选特征以组成候选特征池；以及(D)对候选特征池中的各个候选特征进行重要性的再排序，并根据再排序结果从候选特征池中选择重要性较高的至少一个候选特征作为重要特征。

[0009] 可选地，在所述方法中，在步骤(C)中，基于第一数量的历史数据记录进行预排序；在步骤(D)中，基于第二数量的历史数据记录进行再排序，并且，第二数量不少于第一数量。

[0010] 可选地，在所述方法中，第二数量的历史数据记录包括第一数量的历史数据记录。

[0011] 可选地，在所述方法中，在步骤(C)中，根据预排序结果从所述至少一个候选特征中筛选出重要性较高的候选特征以组成候选特征池。

[0012] 可选地，在所述方法中，在步骤(C)中，通过以下处理来进行预排序：针对每一个候选特征，得到预排序单特征机器学习模型，基于各个预排序单特征机器学习模型的效果来确定各个候选特征的重要性，其中，预排序单特征机器学习模型对应所述每一个候选特征。

[0013] 可选地，在所述方法中，在步骤(C)中，通过以下处理来进行预排序：针对每一个候选特征，得到预排序整体机器学习模型，基于各个预排序整体机器学习模型的效果来确定各个候选特征的重要性，其中，预排序整体机器学习模型对应预排序基本特征子集和所述每一个候选特征。

[0014] 可选地，在所述方法中，在步骤(C)中，通过以下处理来进行预排序：针对每一个候选特征，得到预排序复合机器学习模型，基于各个预排序复合机器学习模型的效果来确定各个候选特征的重要性，其中，预排序复合机器学习模型包括基于提升框架的预排序基本子模型和预排序附加子模型，其中，预排序基本子模型对应预排序基本特征子集，预排序附加子模型对应所述每一个候选特征。

[0015] 可选地，在所述方法中，预排序基本特征子集包括由所述多个属性信息之中的至少一个属性信息自身单独表示的单位特征，并且，候选特征包括由所述单位特征组合而成的组合特征。

[0016] 可选地，在所述方法中，在步骤(D)中，通过以下处理来进行再排序：针对候选特征池中的每一个候选特征，得到再排序单特征机器学习模型，基于各个再排序单特征机器学习模型的效果来确定各个候选特征的重要性，其中，再排序单特征机器学习模型对应所述每一个候选特征。

[0017] 可选地，在所述方法中，在步骤(D)中，通过以下处理来进行再排序：针对候选特征池中的每一个候选特征，得到再排序整体机器学习模型，基于各个再排序整体机器学习模型的效果来确定各个候选特征的重要性，其中，再排序复合机器学习模型对应再排序基本特征子集和所述每一个候选特征。

[0018] 可选地，在所述方法中，在步骤(D)中，通过以下处理来进行再排序：针对候选特征池中的每一个候选特征，得到再排序复合机器学习模型，基于各个再排序复合机器学习模型的效果来确定各个候选特征的重要性，其中，再排序复合机器学习模型包括基于提升框架的再排序基本子模型和再排序附加子模型，其中，再排序基本子模型对应再排序基本特征子集，再排序附加子模型对应所述每一个候选特征。

[0019] 可选地，在所述方法中，再排序基本特征子集包括由所述多个属性信息之中的至少一个属性信息自身单独表示的单位特征，并且，候选特征包括由所述单位特征组合而成的组合特征。

[0020] 可选地，所述方法还包括：(E)检验所述重要特征是否适于作为机器学习样本的特征。

[0021] 可选地，在所述方法中，在步骤(E)中，利用基于由所述多个属性信息之中的至少一个属性信息自身单独表示的单位特征的机器学习模型在引入所述重要特征之后的效果变化来检验所述重要特征是否适于作为机器学习样本的特征。

[0022] 可选地，在所述方法中，在检验结果为所述重要特征不适于作为机器学习样本的特征的情况下，根据预排序结果从所述至少一个候选特征中筛选出另外的一部分候选特征以组成新的候选特征池，并重新执行步骤(D)和步骤(E)。

[0023] 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重要特征的计算机可读介质，其中，在所述计算机可读介质上记录有用于执行如上所述的方法的计算机程序。

[0024] 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重要特征的计算装置，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，执行如上所述的方法。

[0025] 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重要特征的系统，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多个属性信息；候选特征生成装置，用于基于所述多个属性信息生成至少一个候选特征；预排序装置，用于对所述至少一个候选特征进行重要性的预排序，并根据预排序结果从所述至少一个候选特征中筛选出一部分候选特征以组成候选特征池；以及再排序装置，用于对候选特征池中的各个候选特征进行重要性的再排序，并根据再排序结果从候选特征池中选择重要性较高的至少一个候选特征作为重要特征。

[0026] 可选地，在所述系统中，预排序装置基于第一数量的历史数据记录进行预排序；再排序装置基于第二数量的历史数据记录进行再排序，并且，第二数量不少于第一数量。

[0027] 可选地，在所述系统中，第二数量的历史数据记录包括第一数量的历史数据记录。

[0028] 可选地，在所述系统中，预排序装置根据预排序结果从所述至少一个候选特征中筛选出重要性较高的候选特征以组成候选特征池。

[0029] 可选地，在所述系统中，预排序装置通过以下处理来进行预排序：针对每一个候选特征，得到预排序单特征机器学习模型，基于各个预排序单特征机器学习模型的效果来确定各个候选特征的重要性，其中，预排序单特征机器学习模型对应所述每一个候选特征。

[0030] 可选地，在所述系统中，预排序装置通过以下处理来进行预排序：针对每一个候选特征，得到预排序整体机器学习模型，基于各个预排序整体机器学习模型的效果来确定各个候选特征的重要性，其中，预排序整体机器学习模型对应预排序基本特征子集和所述每一个候选特征。

[0031] 可选地，在所述系统中，预排序装置通过以下处理来进行预排序：针对每一个候选特征，得到预排序复合机器学习模型，基于各个预排序复合机器学习模型的效果来确定各个候选特征的重要性，其中，预排序复合机器学习模型包括基于提升框架的预排序基本子模型和预排序附加子模型，其中，预排序基本子模型对应预排序基本特征子集，预排序附加子模型对应所述每一个候选特征。

[0032] 可选地，在所述系统中，预排序基本特征子集包括由所述多个属性信息之中的至少一个属性信息自身单独表示的单位特征，并且，候选特征包括由所述单位特征组合而成的组合特征。

[0033] 可选地，在所述系统中，再排序装置通过以下处理来进行再排序：针对候选特征池中的每一个候选特征，得到再排序单特征机器学习模型，基于各个再排序单特征机器学习模型的效果来确定各个候选特征的重要性，其中，再排序单特征机器学习模型对应所述每

一个候选特征。

[0034] 可选地，在所述系统中，再排序装置通过以下处理来进行再排序：针对候选特征池中的每一个候选特征，得到再排序整体机器学习模型，基于各个再排序整体机器学习模型的效果来确定各个候选特征的重要性，其中，再排序复合机器学习模型对应再排序基本特征子集和所述每一个候选特征。

[0035] 可选地，在所述系统中，再排序装置通过以下处理来进行再排序：针对候选特征池中的每一个候选特征，得到再排序复合机器学习模型，基于各个再排序复合机器学习模型的效果来确定各个候选特征的重要性，其中，再排序复合机器学习模型包括基于提升框架的再排序基本子模型和再排序附加子模型，其中，再排序基本子模型对应再排序基本特征子集，再排序附加子模型对应所述每一个候选特征。

[0036] 可选地，在所述系统中，再排序基本特征子集包括由所述多个属性信息之中的至少一个属性信息自身单独表示的单位特征，并且，候选特征包括由所述单位特征组合而成的组合特征。

[0037] 可选地，所述系统还包括：检验装置，用于检验所述重要特征是否适于作为机器学习样本的特征。

[0038] 可选地，在所述系统中，检验装置利用基于由所述多个属性信息之中的至少一个属性信息自身单独表示的单位特征的机器学习模型在引入所述重要特征之后的效果变化来检验所述重要特征是否适于作为机器学习样本的特征。

[0039] 可选地，在所述系统中，在检验结果为所述重要特征不适于作为机器学习样本的特征的情况下，预排序装置根据预排序结果从所述至少一个候选特征中筛选出另外的一部分候选特征以组成新的候选特征池，以便再排序装置和检验装置重新执行相应的操作。

[0040] 在根据本发明示例性实施例的用于确定机器学习样本的重要特征的方法及系统中，通过特定方式的预排序和再排序从候选特征中筛选出相对重要的特征，从而可在使用较少运算资源的情况下有效地确定重要特征，有助于提升机器学习模型的效果。

附图说明

[0041] 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：

[0042] 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重要特征的系统的框图；

[0043] 图2示出根据本发明另一示例性实施例的用于确定机器学习样本的重要特征的系统的框图；

[0044] 图3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重要特征的方法的流程图；以及

[0045] 图4示出根据本发明另一示例性实施例的用于确定机器学习样本的重要特征的方法的流程图。

具体实施方式

[0046] 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。

[0047] 在本发明的示例性实施例中，通过以下方式来确定重要特征：基于数据记录的属性信息来生成用于执行机器学习的候选特征，通过预排序从中筛选出一部分候选特征，进而通过再排序从筛选出的候选特征中选取重要性较高的特征。

[0048] 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，本发明的示例性实施例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。

[0049] 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重要特征的系统的框图。图1所示的系统包括数据记录获取装置100、候选特征生成装置200、预排序装置300和再排序装置400。

[0050] 具体说来，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多个属性信息。这里，作为示例，数据记录获取装置100可获取已经标记过的历史数据记录，以用于进行有监督机器学习。

[0051] 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接收的数据。这些数据可涉及个人、企业或组织的属性信息，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属性信息，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。

[0052] 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/测试过程。这些数据可来源于期望获取模型预测结果的实体内部，例如，来源于期望获取预测结果的银行、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。

[0053] 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。

[0054] 候选特征生成装置200用于基于所述多个属性信息生成至少一个候选特征。这里，候选特征生成装置200可根据任何适当的特征处理方式，通过对所述属性信息进行处理来生成候选特征。

[0055] 具体说来，针对历史数据记录的至少一部分属性信息，可产生相应的连续特征，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。

[0056] 举例说来，可将历史数据记录中的某种连续值属性信息直接作为对应的连续特征，例如，可将距离、年龄、金额等属性信息直接作为相应的连续特征。也就是说，所述每一个连续特征可由所述多个属性信息之中的连续值属性信息自身形成。或者，也可通过对历史数据记录中的某些属性信息(例如，连续值属性和/或离散值属性信息)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。特别地，所述连续特征可通过对所述多个属性信息之中的离散值属性信息进行连续变换而形成。作为示例，所述连续变换可指示对所述离散值属性信息的取值进行统计。例如，连续特征可指示某些离散值属性信息关于机器学习模型的预测目标的统计信息。举例说来，在预测购买概率的示例中，可将卖家商户编号这一离散值属性信息变换为关于相应卖家商户编码的历史购买行为的概率统计特征。

[0057] 除了连续特征之外，候选特征生成装置200还可产生离散特征。类似地，可将历史数据记录中的某种离散值属性信息直接作为对应的离散特征，或者，也可通过对历史数据记录中的某些属性信息(例如，连续值属性和/或离散值属性信息)进行处理，以得到相应的离散特征。

[0058] 作为示例，在此过程中，候选特征生成装置200可根据需要对连续值属性信息进行离散化和/或对离散值属性信息进行连续化等，并且，候选特征生成装置200可对原始或经过处理的不同属性值信息进行进一步的运算或结合等。甚至，特征之间也可以进行任意组合或运算，例如，离散特征之间可进行笛卡尔积组合。

[0059] 作为可选方式，为了进一步处理连续特征，候选特征生成装置200可执行至少一种分箱运算，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离散特征。

[0060] 这里，分箱(binning)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，由候选特征生成装置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。

[0061] 候选特征生成装置200执行的分箱运算可以在分箱方式和/或分箱参数方面存在差异。例如，所述至少一种分箱运算可以是种类相同但具有不同运算参数(例如，深度、宽度等)的分箱运算，也可以是不同种类的分箱运算。相应地，每一种分箱运算可得到一个分箱特征，这些分箱特征共同组成一个分箱组特征，该分箱组特征可体现出不同分箱运算，从而提升了机器学习素材的有效性，为机器学习模型的训练/预测提供了较好的基础。

[0062] 应注意，本发明的示例性实施例并不限制生成候选特征的具体方式，任何经由诸如特征工程的处理所得到的特征均可作为候选特征。

[0063] 预排序装置300用于对所述至少一个候选特征进行重要性的预排序，并根据预排序结果从所述至少一个候选特征中筛选出一部分候选特征以组成候选特征池。

[0064] 作为示例，预排序装置300可利用任何判断特征重要性的手段来衡量各个候选特征的重要性。通过预排序，能够获知所述各个候选特征的重要性顺序，在此基础上，预排序装置300可从中筛选出一部分候选特征以组成候选特征池，这里，筛选出的候选特征可表现为在预测作用方面的某种一致性，使得可仅从中筛选出重要性较高(即，预测力较强)的特征以作为机器学习样本的重要特征。

[0065] 例如，预排序装置300可筛选重要性较高的一部分候选特征(例如，100个候选特征中，可筛选出第一重要到第十重要的特征)、重要性成间隔排列的一部分候选特征(例如，100个候选特征中，可筛选出第一重要的特征、第十一重要的特征、第二十一重要的特征…、第九十一重要的特征)等。筛选出的候选特征可构成候选特征池，以便从中进一步筛选出重要性较高的特征。

[0066] 相应地，再排序装置400用于对候选特征池中的各个候选特征进行重要性的再排序，并根据再排序结果从候选特征池中选择重要性较高的至少一个候选特征作为重要特征。这里，再排序装置400可利用任何判断特征重要性的手段来衡量候选特征池中的各个候选特征的重要性，例如，再排序装置400可采用与预排序装置300相同的方式来衡量候选特征的重要性，只是在判断时基于数量更多和/或质量更高的数据记录以作出更为准确的判断。再排序装置400可选择候选特征池中最重要的预定数量个候选特征作为重要特征，这里，作为示例，重要特征可直接作为机器学习样本的特征，或可对重要特征进行进一步的验证以决定是否将其作为机器学习样本的特征。作为示例，如果在当前候选特征池中没有筛选到合适的特征，则可根据预排序结果重新确定新的候选特征池，例如，在100个候选特征中，可再次筛选出第十一重要到第二十重要的特征；或者，在100个候选特征中，可再次筛选出第二重要的特征、第十二重要的特征、第二十二重要的特征…、第九十二重要的特征。

[0067] 在图1所示的系统中，预排序装置300和再排序装置400均涉及对特征重要性的确定，相应地，作为可选方式，上述两个装置可共享一部分运算参数或结果，以节省资源。

[0068] 应注意：图1所示的各个装置可被配置为由软件、硬件和/或固件组成的各个单元，这些单元中的某些单元或全部单元可被集成为一体或共同协作以完成特定功能。

[0069] 图2示出根据本发明另一示例性实施例的用于确定机器学习样本的重要特征的系统的框图。在图2所示的系统中，除了数据记录获取装置100、候选特征生成装置200、预排序装置300和再排序装置400之外，还进一步包括检验装置500，用于检验所述重要特征是否适于作为机器学习样本的特征。

[0070] 这里，数据记录获取装置100、候选特征生成装置200、预排序装置300和再排序装置400可按照参照图1描述的方式来进行操作，这里将不再赘述细节。此外，由再排序装置400每次选择出的特征并不会被直接用作机器学习样本的重要特征，而是需经过检验装置500的验证处理。作为示例，检验装置500可通过将选择的重要特征融入将针对预测问题执行预测的实际机器学习模型来检验其是否适于作为机器学习样本的特征，例如，检验装置500可将待验证的重要特征引入基于已经验证过的特征的机器学习模型，并通过衡量模型

的效果变化来检验所述待验证的重要特征是否适于作为机器学习样本的特征。

[0071] 图1和图2所示的系统旨在产生机器学习样本的重要特征，该系统可独立存在，这里，应注意，所述系统获取数据记录的方式并不受限制，也就是说，作为示例，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。此外，上述系统也可集成到模型训练系统中，作为完成特征处理的组成部分。

[0072] 在根据本发明的模型训练系统中，除了数据记录获取装置100、候选特征生成装置200、预排序装置300和再排序装置400之外，还包括机器学习样本生成装置和机器学习模型训练装置(未示出)。

[0073] 具体说来，数据记录获取装置100、候选特征生成装置200、预排序装置300和再排序装置400可按照在图1所示的方式进行操作，其中，作为示例，数据记录获取装置100可获取已经标记过的历史数据记录。

[0074] 此外，机器学习样本生成装置用于产生至少包括一部分所选择的重要特征的机器学习样本。也就是说，在由机器学习样本生成装置产生的机器学习样本中，包括由再排序装置400筛选出的一部分或全部重要特征，此外，作为可选方式，机器学习样本还可包括基于数据记录的属性信息产生的任意其他特征，例如，通过对数据记录的属性信息进行特征处理而得到的特征等。

[0075] 具体说来，机器学习样本生成装置可产生机器学习训练样本，特别地，作为示例，在有监督学习的情况下，机器学习样本生成装置所产生的机器学习训练样本可包括特征和标记(label)两部分。

[0076] 机器学习模型训练装置用于基于机器学习训练样本来训练机器学习模型。这里，机器学习模型训练装置可采用任何适当的机器学习算法(例如，对数几率回归)，从机器学习训练样本学习出适当的机器学习模型。作为示例，机器学习模型训练装置可采用与预排序装置300或再排序装置400为了衡量相关特征重要性所采用的模型相同或类似的机器学习算法。

[0077] 在上述示例中，可训练出较为稳定且预测效果较好的机器学习模型。

[0078] 以下结合图3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重要特征的方法的流程图。这里，作为示例，图3所示的方法可由图1所示的系统及其装置来执行，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3所示的方法。为了描述方便，假设图3所示的方法由图1所示的系统来执行。

[0079] 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多个属性信息。

[0080] 这里，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数据，或对采集的原始数据进行处理，使得处理后的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。

[0081] 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。

[0082] 上述获取的数据记录可来源于相同或不同的数据源，比如，每条数据记录可以是不同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拼接为完整的数据记录。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。

[0083] 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。

[0084] 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。

[0085] 在获取了历史数据记录之后，在步骤S200中，由候选特征生成装置200基于所述多个属性信息生成至少一个候选特征

[0086] 如上所述，候选特征生成装置200可根据任何适当的特征处理方式，通过对属性信息进行处理来得到相应的候选特征。

[0087] 作为示例，在该过程中，候选特征生成装置200可通过执行多分箱处理来进行连续特征的离散化。具体说来，候选组合特征生成装置200可针对每一个连续特征，执行至少一种分箱运算，以生成由至少一个分箱特征组成的离散特征，其中，每种分箱运算对应一个分箱特征。上述分箱特征组成的离散特征可代替原始的连续特征而参与离散特征之间的自动组合，或者，所述离散特征可再次经历连续变换以得到新的连续特征。

[0088] 这里，候选组合特征生成装置200可按照各种分箱方式和/或分箱参数来执行分箱运算。

[0089] 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。

[0090] 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。这里，分箱特征可指示连续特征按照对应的分箱运算被分到了哪个箱子。也就是说，执行分箱运算以产生与每一个连续特征对应的多维度的分箱特征，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的连续特征对应的分箱特征可表示为[0,0,0,0,0,0,1,0,0,0]。

[0091] 以上示出了通过对连续特征执行分箱运算而得到离散特征的示例，应注意，根据本发明的示例性实施例，还可通过设置分箱特征中相关维度的取值来获得可用作连续特征的分箱特征。具体说来，在通过对连续特征执行分箱运算而得到的多维度的分箱特征中，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。除此之外，还可对分箱特征的取值进行归一化处理，以便于执行运算。假设执行分箱运算的第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。

[0092] 其中，x'_{ij}可如下式表示：

[0093] $$x'_{ij} = (x_{ij} - \min_i) \times \frac{B}{\max_i - \min_i} - \text{BinID},$$

[0094] 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，

[0095] $$\text{BinID} = \left\lfloor (x_{ij} - \min_i) \times \frac{B}{\max_i - \min_i} \right\rfloor$$，其中，⌊ ⌋为向下取整运算符号。

[0096] 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分箱特征(6,0.15)。

[0097] 这里，为了获得上述特征(BinID,x'_{ij})，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，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。

[0098] 作为可选方式，在执行特定分箱运算前，还可以通过去除数据样本中可能的离群点来减少数据记录中的噪音。通过这种方式，能进一步提高利用分箱特征进行机器学习的有效性。

[0099] 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。

[0100] 作为示例，针对连续特征执行的至少一种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，采用的分箱方式相同但划分的粒度不同，这使得产生的分箱特征能够更好地刻画原始数据记录的规律，从而更有利于机器学习模型的训练与预测。特别地，至少一种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，分箱运算可按照值2、值4、值8、值16等的宽度来进行等宽分箱。或者，至少一种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，分箱运算可按照值2、值4、值6、值8等的宽度来进行等宽分箱。

[0101] 作为另一示例，针对连续特征执行的至少一种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，分箱运算采用的分箱方式相同但划分的粒度不同，这使得产生的分箱特征能够更好地刻画原始数据记录的规律，从而更有利于机器学习模型的训练与预测。特别地，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，分箱运算可按照值10、值100、值1000、值10000等的深度来进行等深分箱。或者，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，分箱运算可按照值10、值20、值30、值40等的深度来进行等深分箱。

[0102] 作为示例，针对每一个连续特征，在通过执行分箱运算而得到了相应的至少一个分箱特征之后，可利用其中的一个或多个分箱特征来表示与连续特征对应的特征，该特征可看做相关分箱特征的集合，与连续特征和/或离散特征进行组合。这里，应理解，由于分箱运算的执行而使得连续特征被离散化地置入相应的特定箱中，然而，应注意，根据本发明的示例性实施例，在转换后的分箱特征中，每个维度既可以指示箱子中是否被分配了连续特征的离散值(例如，“0”或“1”)，也可以指示具体的连续数值(例如，特征值、平均值、中间值、边界值、归一化值等)。相应地，在机器学习中具体应用各个维度的离散值(例如，针对分类问题)或连续数值(例如，针对回归问题)时，可进行离散值之间的组合(例如，笛卡尔积等)或连续数值之间的组合(例如，算术运算组合等)。

[0103] 如上所述，根据本发明的示例性实施例，可对连续特征执行至少一种分箱运算。这里，所述至少一种分箱运算可通过任何适当的方式来确定，例如，可借助技术人员或业务人员的经验来确定，也可经由技术手段来自动确定。作为示例，可基于分箱特征的重要性来有效地确定具体的分箱运算方式。

[0104] 接下来，在步骤S300中，由预排序装置300对所述至少一个候选特征进行重要性的预排序，并根据预排序结果从所述至少一个候选特征中筛选出一部分候选特征以组成候选特征池。

[0105] 这里，预排序装置300可利用任何判断特征重要性的手段来衡量各个候选特征的重要性。

[0106] 例如，预排序装置300可针对每一个候选特征，得到预排序单特征机器学习模型，基于各个预排序单特征机器学习模型的效果来确定各个候选特征的重要性，其中，预排序单特征机器学习模型对应所述每一个候选特征。

[0107] 作为示例，假设存在N(N为大于1的整数)个候选特征f_n，其中，n∈[1,N]。相应地，预排序装置300可利用至少一部分历史数据记录来构建N个预排序单特征机器学习模型(其中，每一个预排序单特征机器学习模型基于相应的单个候选特征f_n来针对机器学习问题进行预测)，然后衡量这N个预排序单特征机器学习模型在相同测试数据集上的效果(例如，AUC(ROC(受试者工作特征，Receiver Operating Characteristic)曲线下的面积，Area Under ROC Curve)、MAE(平均绝对误差，Mean Absolute Error)等)，并基于效果的排序来确定各个候选特征的重要性顺序。

[0108] 又例如，预排序装置300可针对每一个候选特征，得到预排序整体机器学习模型，基于各个预排序整体机器学习模型的效果来确定各个候选特征的重要性，其中，预排序整体机器学习模型对应预排序基本特征子集和所述每一个候选特征。作为示例，这里的预排序整体机器学习模型可以是对数几率回归(LR)模型；相应地，预排序整体机器学习模型的样本由预排序基本特征子集和所述每一个候选特征组成。

[0109] 作为示例，假设存在N个候选特征f_n，相应地，预排序装置300可利用至少一部分历史数据记录来构建N个预排序整体机器学习模型(其中，每一个预排序整体机器学习模型的样本特征包括固定的预排序基本特征子集和相应的候选特征f_n)，然后衡量这N个预排序整体机器学习模型在相同测试数据集上的效果(例如，AUC、MAE等)，并基于效果的排序确定各个候选特征的重要性顺序。

[0110] 又例如，预排序装置300可针对每一个候选特征，得到预排序复合机器学习模型，基于各个预排序复合机器学习模型的效果来确定各个候选特征的重要性，其中，预排序复合机器学习模型包括基于提升框架(例如，梯度提升框架)的预排序基本子模型和预排序附加子模型，其中，预排序基本子模型对应预排序基本特征子集，预排序附加子模型对应所述每一个候选特征。

[0111] 作为示例，假设存在N个候选特征f_n，相应地，预排序装置300可利用至少一部分历史数据记录来构建N个预排序复合机器学习模型(其中，每一个预排序复合机器学习模型基于固定的预排序基本特征子集和相应的候选特征f_n，按照提升框架来针对机器学习问题进行预测)，然后衡量这N个预排序复合机器学习模型在相同测试数据集上的效果(例如，AUC、MAE等)，并基于效果的排序确定各个候选特征的重要性顺序。优选地，为了进一步提高运算效率并降低资源消耗，预排序装置300可通过在固定预排序基本子模型的情况下，分别针对每一个候选特征f_n训练预排序附加子模型来构建各个预排序复合机器学习模型。

[0112] 根据本发明的示例性实施例，预排序基本特征子集可固定地应用于所有相关预排序整体机器学习模型或预排序复合机器学习模型中的预排序基本子模型，这里，预排序基本特征子集可包括由所述多个属性信息之中的至少一个属性信息自身单独表示的单位特征，例如，可将历史数据记录的一部分属性信息或全部属性信息直接作为预排序基本特征。在此情况下，作为示例，候选特征可包括由所述单位特征组合而成的组合特征。此外，作为示例，可考虑实际的机器学习问题，基于估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预排序基本特征。

[0113] 在通过预排序确定了各个候选特征的重要性顺序之后，预排序装置300可基于排序结果从候选特征之中筛选出至少一部分以组成候选特征池。如上所述，可优先筛选在预测作用方面具有一致性的重要候选特征来组成候选特征池，以便有效地从中确定重要特

征。例如，预排序装置300可根据预排序结果从所述至少一个候选特征中筛选出重要性较高的候选特征以组成候选特征池。

[0114] 假设存在1000个候选特征，预排序装置300可从中筛选出预排序结果中最为重要的10个候选特征以组成候选特征池。

[0115] 接下来，在步骤S400中，由再排序装置400对候选特征池中的各个候选特征进行重要性的再排序，并根据再排序结果从候选特征池中选择重要性较高的至少一个候选特征作为重要特征。

[0116] 这里，再排序装置400可利用任何判断特征重要性的手段来衡量候选特征池中的各个候选特征的重要性。

[0117] 例如，再排序装置400可针对候选特征池中的每一个候选特征，得到再排序单特征机器学习模型，基于各个再排序单特征机器学习模型的效果来确定各个候选特征的重要性，其中，再排序单特征机器学习模型对应所述每一个候选特征。

[0118] 作为示例，假设候选特征池包括10个候选特征。相应地，再排序装置400可利用至少一部分历史数据记录来构建10个再排序单特征机器学习模型(其中，每一个再排序单特征机器学习模型基于相应的单个候选特征来针对机器学习问题进行预测)，然后衡量这10个再排序单特征机器学习模型在相同测试数据集上的效果(例如，AUC、MAE等)，并基于效果的排序来确定候选特征池之中的各个候选特征的重要性顺序。

[0119] 又例如，再排序装置400可针对候选特征池中的每一个候选特征，得到再排序整体机器学习模型，基于各个再排序整体机器学习模型的效果来确定各个候选特征的重要性，其中，再排序复合机器学习模型对应再排序基本特征子集和所述每一个候选特征。作为示例，这里的再排序整体机器学习模型可以是LR模型；相应地，再排序整体机器学习模型的样本由再排序基本特征子集和所述每一个候选特征组成。

[0120] 作为示例，假设候选特征池包括10个候选特征，相应地，再排序装置400可利用至少一部分历史数据记录来构建10个再排序整体机器学习模型(其中，每一个再排序整体机器学习模型的样本特征包括固定的再排序基本特征子集和相应的候选特征)，然后衡量这10个再排序整体机器学习模型在相同测试数据集上的效果(例如，AUC、MAE等)，并基于效果的排序确定候选特征池之中的各个候选特征的重要性顺序。

[0121] 又例如，再排序装置400可针对候选特征池中的每一个候选特征，得到再排序复合机器学习模型，基于各个再排序复合机器学习模型的效果来确定各个候选特征的重要性，其中，再排序复合机器学习模型包括基于提升框架(例如，梯度提升框架)的再排序基本子模型和再排序附加子模型，其中，再排序基本子模型对应再排序基本特征子集，再排序附加子模型对应所述每一个候选特征。

[0122] 作为示例，假设候选特征池包括10个候选特征，相应地，再排序装置400可利用至少一部分历史数据记录来构建10个再排序复合机器学习模型(其中，每一个再排序复合机器学习模型基于固定的再排序基本特征子集和相应的候选特征，按照提升框架来针对机器学习问题进行预测)，然后衡量这10个再排序复合机器学习模型在相同测试数据集上的效果(例如，AUC、MAE等)，并基于效果的排序确定候选特征池之中的各个候选特征的重要性顺序。优选地，为了进一步提高运算效率并降低资源消耗，再排序装置400可通过在固定再排序基本子模型的情况下，分别针对每一个候选特征训练再排序附加子模型来构建各个再排

序复合机器学习模型。

[0123] 根据本发明的示例性实施例，再排序基本特征子集可固定地应用于所有相关再排序整体机器学习模型或再排序复合机器学习模型中的再排序基本子模型，这里，再排序基本特征子集可包括由所述多个属性信息之中的至少一个属性信息自身单独表示的单位特征，例如，可将历史数据记录的一部分属性信息或全部属性信息直接作为再排序基本特征。在此情况下，作为示例，候选特征可包括由所述单位特征组合而成的组合特征。此外，作为示例，可考虑实际的机器学习问题，基于估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再排序基本特征。作为可选方式，再排序基本特征子集和预排序基本特征子集可具有相同的特征。

[0124] 在通过再排序确定了候选特征池之中的各个候选特征的重要性顺序之后，再排序装置400可基于排序结果从候选特征池中筛选出至少一个较为重要的候选特征以作为重要特征。

[0125] 根据本发明的示例性实施例，可通过共享相同的模型部分来进一步有效地控制运算资源。

[0126] 作为示例，在预排序装置300和再排序装置400分别基于各自的提升框架复合机器学习模型来进行相关特征的重要性排序时，例如，可基于相对较多的历史数据记录(例如，全量历史数据记录)来训练共同的基本子模型部分，该部分可作为固定模型部分而分别用作预排序复合机器学习模型中的预排序基本子模型和再排序复合机器学习模型中的再排序基本子模型。进一步地，在共享基本子模型的情况下，可并行地训练与每个重要性待确定的特征相应的预排序附加子模型和再排序附加子模型，使得仅通过历史数据记录的一次读取操作即可同时训练多个模型。

[0127] 此外，根据本发明的示例性实施例，可通过控制相关模型部分的样本训练集规模、样本训练顺序和/或样本训练集质量来进一步确保特征的效果。

[0128] 作为示例，预排序装置300可基于相对较少的历史数据记录来训练预排序单特征机器学习模型，而再排序装置400可基于相对较多的历史数据记录来训练再排序单特征机器学习模型；或者，预排序装置300可基于相对较少的历史数据记录来训练预排序整体机器学习模型，而再排序装置400可基于相对较多的历史数据记录来训练再排序整体机器学习模型；或者，预排序装置300可基于相对较少的历史数据记录来训练预排序附加子模型，而再排序装置400可基于相对较多的历史数据记录来训练再排序附加子模型。这里，再排序装置400采用的历史数据记录可包含至少一部分预排序装置300采用的历史数据记录，或者，再排序装置400采用的历史数据记录可不包含排序单元220采用的任何历史数据记录。除了样本训练集规模方面的差异之外，预排序装置300可与再排序装置400采用相同的历史数据记录集，而只是两者训练时的顺序不同。由此可见，预排序装置300可基于第一数量的历史数据记录进行预排序，再排序装置400可基于第二数量的历史数据记录进行再排序，并且，第二数量不少于第一数量。此外，预排序装置300还可采用与再排序装置400质量不同的样本训练集，例如，预排序装置300可采用质量较低的样本训练集，而再排序装置400可采用质量较高的样本训练集，这样，即使再排序装置400使用了规模较小的样本训练集，也能够确保再排序相关模型的效果。

[0129] 应注意，本发明的示例性实施例并不受限于此，而是可采用任何方式来分别构建

各自的基本子模型，也可采用任何适当的训练数据集。

[0130] 此外，作为可选方式，还可采用迭代的方式来不断地确定新的重要特征，作为示例，在每一轮迭代中，可增加新的候选特征，也可相应地改变每一轮提升框架下的基本子模型所对应的特征子集，例如，先前选出的重要特征可作为新的基本特征而加入相应的基本特征子集。

[0131] 根据本发明的示例性实施例，为了进一步确保重要特征的有效性，还可进一步对重要特征进行验证。图4示出根据本发明另一示例性实施例的用于确定机器学习样本的重要特征的方法的流程图，在该方法中，还可检验选择的重要特征是否适于作为机器学习样本的特征。

[0132] 参照图4，步骤S100、步骤S200、步骤S3000和步骤S400与图3所示的相应步骤类似，这里将不再赘述细节。

[0133] 此外，在步骤S400中得到重要特征之后，所述方法进行到步骤S500，在步骤S500中，可由检验装置500检验所述重要特征是否适于作为机器学习样本的特征。这里，检验装置500可逐个验证重要特征，也可一次性验证多个重要特征。

[0134] 作为示例，检验装置500可利用基于由历史数据记录的多个属性信息之中的至少一个属性信息自身单独表示的单位特征的机器学习模型在引入所述重要特征之后的效果变化来检验所述重要特征是否适于作为机器学习样本的特征。

[0135] 此外，为另一示例，检验装置500可利用基于先前通过验证的重要特征的机器学习模型在引入所述重要特征之后的效果变化来检验所述重要特征是否适于作为机器学习样本的特征。上述机器学习模型的样本除了包括之前已经通过检验的重要特征之外，也可进一步包括其他特征(例如，单位特征)。

[0136] 上述机器学习模型可与预排序基本特征子模型和/或再排序基本特征子模型基于类似的特征子集，并可基于数量较多和/或质量较高的历史数据记录来训练。可选地，所述机器学习模型并不基于提升框架，由此可更为准确地验证所述选择的重要特征是否真正有助于针对机器学习问题来执行预测。

[0137] 这里，检验装置500可判断上述机器学习模型在引入选择的重要特征之后，模型效果的变动是否符合要求(例如，效果增强满足预期或效果减弱可以接受)。具体说来，检验装置500可判断模型效果是否有所增强(例如，模型效果的增强是否达到预定增强程度)；或者，检验装置500可判断模型效果是否仅稍有减弱(例如，模型效果的减弱是否低于预定减弱程度，在这种情况下，模型效果的减弱可被忽略)。当模型的效果变化符合要求时，可确定所述选择的重要特征适于作为机器学习样本的特征。

[0138] 相应地，在检验结果为所述选择的重要特征适于作为机器学习样本的特征的情况下，可在后续将所述选择的重要特征作为机器学习样本的特征来使用；在检验结果为所述选择的重要特征不适于作为机器学习样本的特征的情况下，预排序装置300可根据预排序结果从所述至少一个候选特征中筛选出另外的一部分候选特征以组成新的候选特征池，以便再排序装置400和检验装置500重新执行相应的操作。

[0139] 图1和图2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，这些装置或单元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模块。此外，这些装置或单元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。

[0140] 以上参照图1到图4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重要特征的系统及其方法。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,例如,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重要特征的计算机可读介质,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)获取历史数据记录,其中,所述历史数据记录包括多个属性信息;(B)基于所述多个属性信息生成至少一个候选特征;(C)对所述至少一个候选特征进行重要性的预排序,并根据预排序结果从所述至少一个候选特征中筛选出一部分候选特征以组成候选特征池;以及(D)对候选特征池中的各个候选特征进行重要性的再排序,并根据再排序结果从候选特征池中选择重要性较高的至少一个候选特征作为重要特征。

[0141] 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。

[0142] 应注意,根据本发明示例性实施例的重要特征确定系统及其相应的机器学习模型训练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,以实现相应的功能。

[0143] 另一方面,图1到图4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。

[0144] 例如,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重要特征的方法。

[0145] 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。

[0146] 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。

[0147] 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。

[0148] 根据本发明示例性实施例的重要特征确定方法以及机器学习模型训练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。

[0149] 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。

[0150] 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。

[0151] 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。

[0152] 根据本发明示例性实施例的重要特征确定方法以及相应的机器学习模型训练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。

[0153] 例如，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重要特征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)获取历史数据记录，其中，所述历史数据记录包括多个属性信息；(B)基于所述多个属性信息生成至少一个候选特征；(C)对所述至少一个候选特征进行重要性的预排序，并根据预排序结果从所述至少一个候选特征中筛选出一部分候选特征以组成候选特征池；以及(D)对候选特征池中的各个候选特征进行重要性的再排序，并根据再排序结果从候选特征池中选择重要性较高的至少一个候选特征作为重要特征。

[0154] 以上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，本发明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。

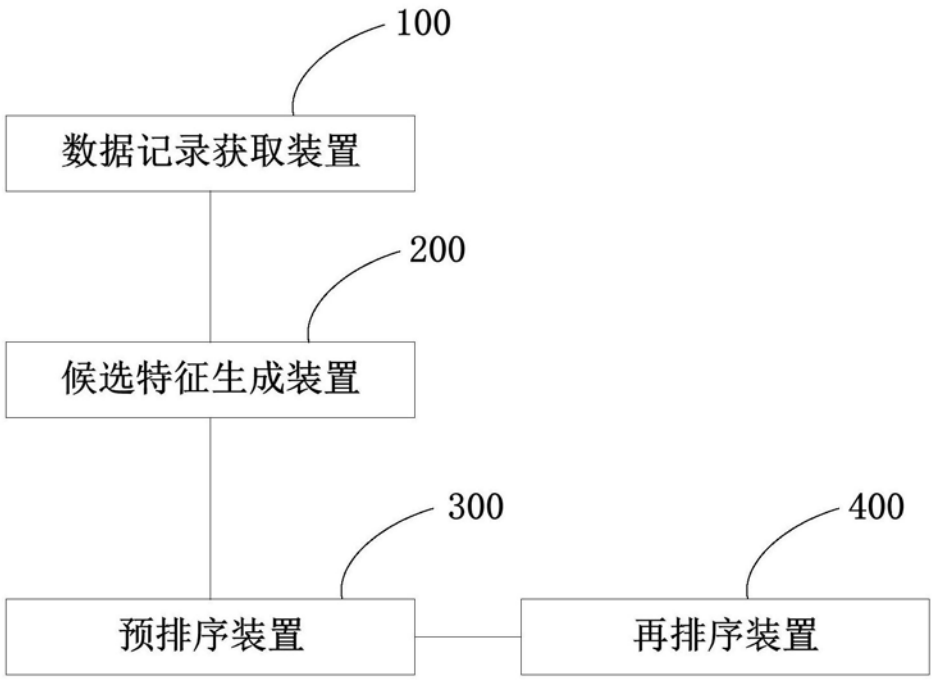

100
数据记录获取装置
200
候选特征生成装置
300
预排序装置
400
再排序装置

图1

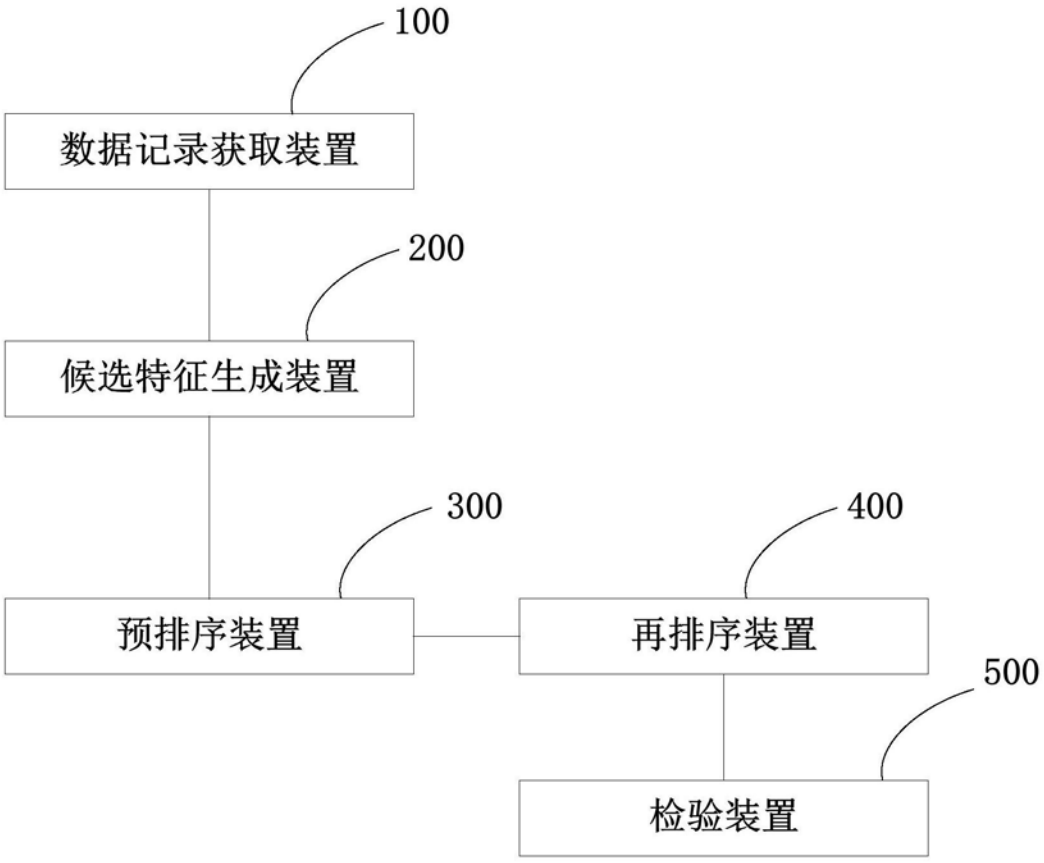

100
数据记录获取装置
200
候选特征生成装置
300
预排序装置
400
再排序装置
500
检验装置

图2

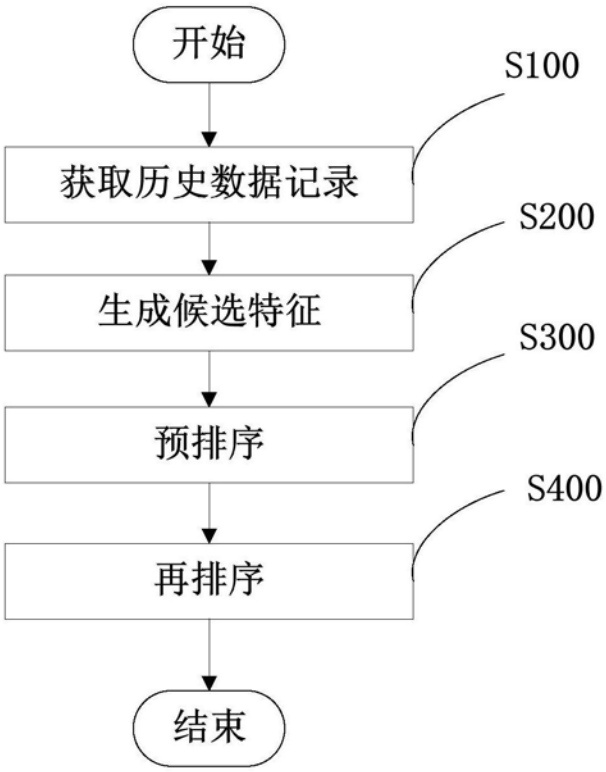


图3

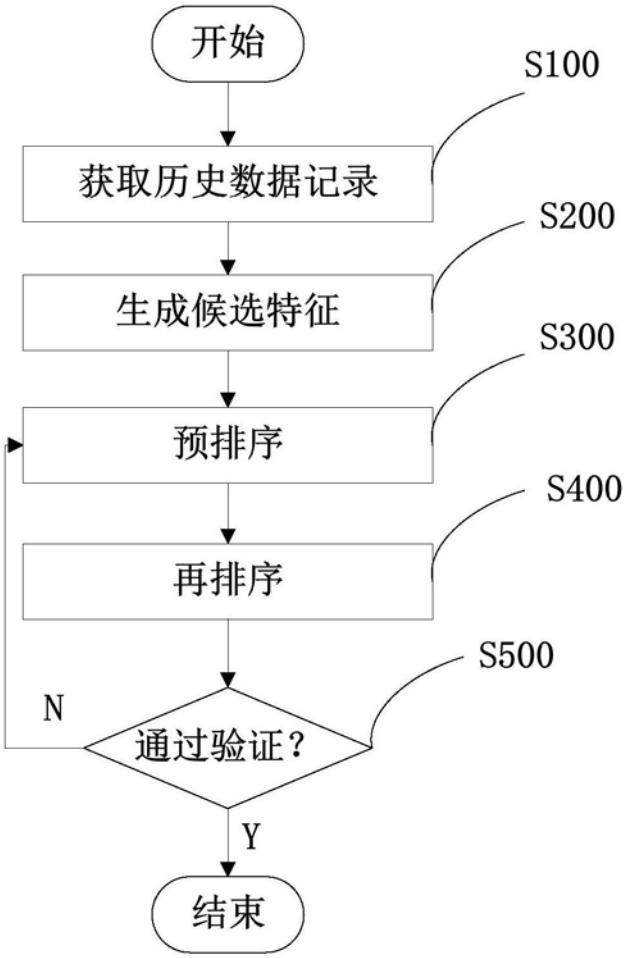

开始
S100
获取历史数据记录
S200
生成候选特征
S300
预排序
S400
再排序
S500
N
通过验证？
Y
结束

图4