(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 112328657 A
(43)申请公布日 2021.02.05

(21)申请号 202011211554.8

(22)申请日 2020.11.03

(71)申请人 中国平安人寿保险股份有限公司
地址 518000 广东省深圳市福田区益田路5033号平安金融中心14、15、16、37、41、44、45、46层

(72)发明人 刘波

(74)专利代理机构 深圳市世联合知识产权代理有限公司 44385
代理人 汪琳琳

(51)Int.Cl.
G06F 16/2458(2019.01)
G06F 16/28(2019.01)
G06F 16/9535(2019.01)
G06K 9/62(2006.01)

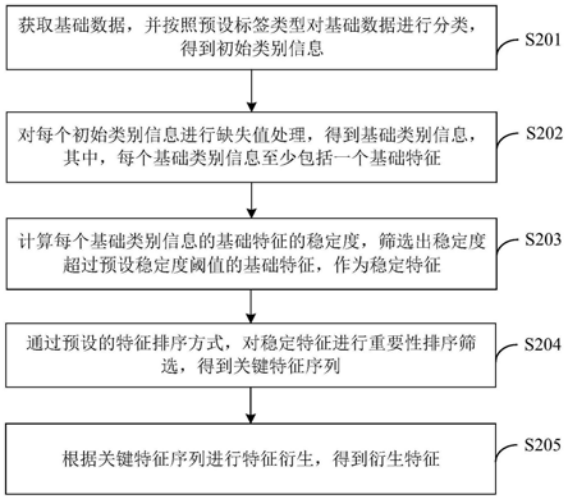
权利要求书2页 说明书11页 附图3页

(54)发明名称

特征衍生方法、装置、计算机设备及介质

(57)摘要

本发明涉及数据处理领域，公开了一种特征衍生方法、装置、计算机设备及介质，所述方法包括：通过获取基础数据，并按照预设标签类型对基础数据进行分类，得到初始类别信息，进而对每个初始类别信息进行缺失值处理，得到基础类别信息，再计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征，通过预设的特征排序方式，对稳定特征进行重要性排序筛选，得到关键特征序列，再根据关键特征序列进行特征衍生，得到衍生特征。本发明提高了特征衍生的效率。



CN 112328657 A

1.一种特征衍生方法,其特征在于,包括:

获取基础数据,并按照预设标签类型对所述基础数据进行分类,得到初始类别信息;

对每个所述初始类别信息进行缺失值处理,得到基础类别信息,其中,每个基础类别信息至少包括一个基础特征;

计算每个基础类别信息的基础特征的稳定度,筛选出稳定度超过预设稳定度阈值的基础特征,作为稳定特征;

通过预设的特征排序方式,对所述稳定特征进行重要性排序筛选,得到关键特征序列;

根据所述关键特征序列进行特征衍生,得到衍生特征。

2.如权利要求1所述的特征衍生方法,其特征在于,所述对每个所述初始类别信息进行缺失值处理,得到基础类别信息包括:

针对每个初始类别信息,获取所述初始类别信息中每个基础特征对应的特征值;

对所述特征值进行数据校验,将未通过校验的特征值作为缺失值;

对每个基础特征对应的缺失值进行统计,并将缺失值与所有特征值的比例超过预设比例的基础特征,作为无效特征,并从所述初始类别信息中移除所述无效特征,得到基础类别信息。

3.如权利要求1所述的特征衍生方法,其特征在于,所述计算每个基础类别信息的基础特征的稳定度,筛选出稳定度超过预设稳定度阈值的基础特征,作为稳定特征包括:

计算每个基础特征的信息值IV,并根据所述信息值IV进行特征筛选,得到关键特征;

通过预设方式,计算所述关键特征的稳定度指标PSI,将所述稳定度指标PSI超过预设稳定度阈值的关键特征,作为稳定特征。

4.如权利要求3所述的特征衍生方法,其特征在于,所述基础特征包括连续型的特征,所述计算每个基础特征的信息值IV包括:

针对对基础特征中数据类型为连续型的特征,进行分箱处理,将连续型的特征转化为离散型特征;

针对所有离散型特征进行独热编码,得到数字化变量;

根据数字化变量,计算每个特征对应的信息值IV。

5.如权利要求1所述的特征衍生方法,其特征在于,所述特征衍生的方式包括特征组合、特征交叉、图像特征生成和文本特征生成中的至少一项。

6.如权利要求1至5所述的特征衍生方法,其特征在于,在所述根据所述关键特征序列进行特征衍生,得到衍生特征之后,所述特征衍生方法还包括:

对所述衍生特征进行稳定性校验;

若所述衍生特征的稳定度超过预设稳定度阈值,保留所述衍生特征,并将所述衍生特征作为所述基础特征,否则,剔除所述衍生特征。

7.一种特征衍生装置,其特征在于,包括:

信息获取模块,用于获取基础数据,并按照预设标签类型对所述基础数据进行分类,得到初始类别信息;

数据处理模块,用于对每个所述初始类别信息进行缺失值处理,得到基础类别信息,其中,每个基础类别信息至少包括一个基础特征;

特征筛选模块,用于计算每个基础类别信息的基础特征的稳定度,筛选出稳定度超过

预设稳定度阈值的基础特征，作为稳定特征；

特征排序模块，用于通过预设的特征排序方式，对所述稳定特征进行重要性排序筛选，得到关键特征序列；

特征衍生模块，用于根据所述关键特征序列进行特征衍生，得到衍生特征。

8.如权利要求7所述的特征衍生装置，其特征在于，所述数据处理模块包括：

特征值确定单元，用于针对每个初始类别信息，获取所述初始类别信息中每个基础特征对应的特征值；

数据校验单元，用于对所述特征值进行数据校验，将未通过校验的特征值作为缺失值；

基础类别信息确定单元，用于对每个基础特征对应的缺失值进行统计，并将缺失值与所有特征值的比例超过预设比例的基础特征，作为无效特征，并从所述初始类别信息中移除所述无效特征，得到基础类别信息。

9.一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求1至6任一项所述的特征衍生方法。

10.一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至6任一项所述的特征衍生方法。

特征衍生方法、装置、计算机设备及介质

技术领域

[0001] 本发明涉及数据处理领域，尤其涉及一种特征衍生方法、装置、计算机设备及介质。

背景技术

[0002] 随着人工智能技术的发展，依据现有数据中的特征进行机器学习建立相关模型已非常常见。例如，在第三方支付平台或网络购物平台的风险防控领域，常依据现有的包含风险特征(例如，交易事件数据中的交易金额、交易频率等特征)的数据进行机器学习得到风控模型，以及，对于一些电信诈骗的风险识别模型等。

[0003] 针对各种风控模型，不法分子会不断的改进作案手段以避开风险防控，使得风险形式不断发生变化，这就需要不断地对风控模型进行改进，以对未来可能出现的新风险做出有效的防控。然而，现有数据中的风险特征无法代表未来的情况，未来的包含新风险特征的数据还没有产生，因此，需要对现有数据中的风险特征进行学习，衍生得到能够反映未来风险的新风险特征，以对风控模型进行改进。其中，对现有特征进行学习衍生得到新特征的过程叫特征衍生。

[0004] 目前，要么依据人工经验进行特征衍生，要么利用穷举的方式进行特征衍生。前者依赖于领域内的专家经验，耗时长、衍生过程慢；后者需要花费大量的计算资源进行计算，耗时也较长、衍生过程也慢。因而，亟需一种高效的特征衍生方法。

发明内容

[0005] 本发明实施例提供一种特征衍生方法、装置、计算机设备和存储介质，以提高特征衍生的效率。

[0006] 为了解决上述技术问题，本申请实施例提供一种特征衍生方法，包括：

[0007] 获取基础数据，并按照预设标签类型对所述基础数据进行分类，得到初始类别信息；

[0008] 对每个所述初始类别信息进行缺失值处理，得到基础类别信息，其中，每个基础类别信息至少包括一个基础特征；

[0009] 计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征；

[0010] 通过预设的特征排序方式，对所述稳定特征进行重要性排序筛选，得到关键特征序列；

[0011] 根据所述关键特征序列进行特征衍生，得到衍生特征。

[0012] 可选地，所述对每个所述初始类别信息进行缺失值处理，得到基础类别信息包括：

[0013] 针对每个初始类别信息，获取所述初始类别信息中每个基础特征对应的特征值；

[0014] 对所述特征值进行数据校验，将未通过校验的特征值作为缺失值；

[0015] 对每个基础特征对应的缺失值进行统计，并将缺失值与所有特征值的比例超过预

设比例的基础特征，作为无效特征，并从所述初始类别信息中移除所述无效特征，得到基础类别信息。

[0016] 可选地，所述计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征包括：

[0017] 计算每个基础特征的信息值IV，并根据所述信息值IV进行特征筛选，得到关键特征；

[0018] 通过预设方式，计算所述关键特征的稳定度指标PSI，将所述稳定度指标PSI超过预设稳定度阈值的关键特征，作为稳定特征。

[0019] 可选地，所述基础特征包括连续型的特征，所述计算每个基础特征的信息值IV包括：

[0020] 针对对基础特征中数据类型为连续型的特征，进行分箱处理，将连续型的特征转化为离散型特征；

[0021] 针对所有离散型特征进行独热编码，得到数字化变量；

[0022] 根据数字化变量，计算每个特征对应的信息值IV。

[0023] 可选地，所述特征衍生的方式包括特征组合、特征交叉、图像特征生成和文本特征生成中的至少一项。

[0024] 可选地，在所述根据所述关键特征序列进行特征衍生，得到衍生特征之后，所述特征衍生方法还包括：

[0025] 对所述衍生特征进行稳定性校验；

[0026] 若所述衍生特征的稳定度超过预设稳定度阈值，保留所述衍生特征，并将所述衍生特征作为所述基础特征，否则，剔除所述衍生特征。

[0027] 为了解决上述技术问题，本申请实施例还提供一种特征衍生装置，包括：

[0028] 信息获取模块，用于获取基础数据，并按照预设标签类型对所述基础数据进行分类，得到初始类别信息；

[0029] 数据处理模块，用于对每个所述初始类别信息进行缺失值处理，得到基础类别信息，其中，每个基础类别信息至少包括一个基础特征；

[0030] 特征筛选模块，用于计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征；

[0031] 特征排序模块，用于通过预设的特征排序方式，对所述稳定特征进行重要性排序筛选，得到关键特征序列；

[0032] 特征衍生模块，用于根据所述关键特征序列进行特征衍生，得到衍生特征。

[0033] 可选地，所述数据处理模块包括：

[0034] 特征值确定单元，用于针对每个初始类别信息，获取所述初始类别信息中每个基础特征对应的特征值；

[0035] 数据校验单元，用于对所述特征值进行数据校验，将未通过校验的特征值作为缺失值；

[0036] 基础类别信息确定单元，用于对每个基础特征对应的缺失值进行统计，并将缺失值与所有特征值的比例超过预设比例的基础特征，作为无效特征，并从所述初始类别信息中移除所述无效特征，得到基础类别信息。

[0037] 可选地，所述特征筛选模块包括：

[0038] 信息值计算单元，用于计算每个基础特征的信息值IV，并根据所述信息值IV进行特征筛选，得到关键特征；

[0039] 稳定度计算单元，用于通过预设方式，计算所述关键特征的稳定度指标PSI，将所述稳定度指标PSI超过预设稳定度阈值的关键特征，作为稳定特征。

[0040] 可选地，基础特征包括连续型的特征，所述信息值计算单元•包括：

[0041] 离散化子单元，用于针对对基础特征中数据类型为连续型的特征，进行分箱处理，将连续型的特征转化为离散型特征；

[0042] 数字化子单元，用于针对所有离散型特征进行独热编码，得到数字化变量；

[0043] 计算子单元，用于根据数字化变量，计算每个特征对应的信息值IV。

[0044] 可选地，所述特征衍生装置还包括：

[0045] 稳定度校验模块，用于对所述衍生特征进行稳定性校验；

[0046] 特征甑选模块，用于若所述衍生特征的稳定度超过预设稳定度阈值，保留所述衍生特征，并将所述衍生特征作为所述基础特征，否则，剔除所述衍生特征。

[0047] 为了解决上述技术问题，本申请实施例还提供一种计算机设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现上述特征衍生方法的步骤。

[0048] 为了解决上述技术问题，本申请实施例还提供一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述特征衍生方法的步骤。

[0049] 本发明实施例提供的特征衍生方法、装置、计算机设备及介质，通过获取基础数据，并按照预设标签类型对基础数据进行分类，得到初始类别信息，进而对每个初始类别信息进行缺失值处理，得到基础类别信息，再计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征，确保后续进行特征衍生的特征的质量，有利于提高后续特征衍生的效率，进而通过预设的特征排序方式，对稳定特征进行重要性排序筛选，得到关键特征序列，再根据关键特征序列进行特征衍生，得到衍生特征。实现从依据重要性进行排序的稳定特征序列中，快速进行特征衍生，提高了特征衍生的效率，以及得到的衍生特征的质量。

附图说明

[0050] 为了更清楚地说明本发明实施例的技术方案，下面将对本发明实施例的描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

[0051] 图1是本申请可以应用于其中的示例性系统架构图；

[0052] 图2是本申请的特征衍生方法的一个实施例的流程图；

[0053] 图3是根据本申请的特征衍生装置的一个实施例的结构示意图；

[0054] 图4是根据本申请的计算机设备的一个实施例的结构示意图。

具体实施方式

[0055] 除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同；本文中在申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请；本申请的说明书和权利要求书及上述附图说明中的术语“包括”和“具有”以及它们的任何变形，意图在于覆盖不排他的包含。本申请的说明书和权利要求书或上述附图中的术语“第一”、“第二”等是用于区别不同对象，而不是用于描述特定顺序。

[0056] 在本文中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

[0057] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通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，都属于本发明保护的范围。

[0058] 请参阅图1，如图1所示，系统架构100可以包括终端设备101、102、103，网络104和服务器105。网络104用以在终端设备101、102、103和服务器105之间提供通信链路的介质。网络104可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

[0059] 用户可以使用终端设备101、102、103通过网络104与服务器105交互，以接收或发送消息等。

[0060] 终端设备101、102、103可以是具有显示屏并且支持网页浏览的各种电子设备，包括但不限于智能手机、平板电脑、电子书阅读器、MP3播放器(Moving Picture E界面显示perts Group Audio Layer III，动态影像专家压缩标准音频层面3)、MP4(Moving Picture E界面显示perts Group Audio Layer IV，动态影像专家压缩标准音频层面4)播放器、膝上型便携计算机和台式计算机等等。

[0061] 服务器105可以是提供各种服务的服务器，例如对终端设备101、102、103上显示的页面提供支持的后台服务器。

[0062] 需要说明的是，本申请实施例所提供的特征衍生方法由服务器执行，相应地，特征衍生装置设置于服务器中。

[0063] 应该理解，图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器，本申请实施例中的终端设备101、102、103具体可以对应的是实际生产中的应用系统。

[0064] 请参阅图2，图2示出本发明实施例提供的一种特征衍生方法，以该方法应用在图1中的服务端为例进行说明，详述如下：

[0065] S201：获取基础数据，并按照预设标签类型对基础数据进行分类，得到初始类别信息。

[0066] 具体地，通过数据采集的方式，获取基础数据，并根据预设标签类型，对基础数据进行分类，得到初始类别信息。

[0067] 其中，数据采集的方式具体包括但不限于：网络爬虫爬取、基于大数据的分布式数

据采集和数据库读取等，具体可依据实际情况进行择取，此处不作限制。

[0068] 其中，预设的标签类型可根据实际情况进行设定，例如，在一具体实施方式中，预设的标签类型包括基础属性、浏览属性、消费属性和信用属性等。

[0069] 需要说明的是，在本实施例中，每个初始类别信息可以包括一个或多个基础特征，例如，在上述示例中，基础属性可以包括年龄、性别、身高、体重以及籍贯等基础特征，浏览属性包括浏览、收藏、转发、评论及点赞等基础特征。

[0070] S202：对每个初始类别信息进行缺失值处理，得到基础类别信息，其中，每个基础类别信息至少包括一个基础特征。

[0071] 具体地，在对获取到的数据进行分类后，需要对数据进行数据预处理，以确保数据质量，考虑到存在数据来源不一致、未及时更新等原因导致的部分数据缺失问题，在本实施例中，先对每个初始类别信息进行缺失值的处理，得到基础类别信息。

[0072] 其中，缺失值是指不符合规范的数值，针对缺失值处理，具体包括但不限于：删除处理、补全处理和更新处理等。具体处理过程也可参考后续实施例的描述，此处不做具体限制。

[0073] S203：计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征。

[0074] 具体地，每个基础类别信息包括一个或多个基础特征，在得到每个基础类别信息后，需要对这些基础特征的稳定性进行评估，保留稳定性好的基础特征，作为稳定特征，在后续通过稳定特征进行特征衍生，得到质量较好的衍生特征，有利于提高后续生成衍生特征的质量。

[0075] 其中，预设稳定度阈值可以根据实际情况进行设定，此处不做限定，在本实施例中，优选值为0.25。

[0076] 生成稳定特征的具体过程，可参考后续实施例的描述，为避免重复，此处不作赘述。

[0077] S204：通过预设的特征排序方式，对稳定特征进行重要性排序筛选，得到关键特征序列。

[0078] 具体地，通过预设的特征排序方式，从多个维度对稳定性特征进行重要性排序，得到关键特征序列，以便后续再进行特征衍生处理时，优先选取排序靠前的特征进行特征衍生。

[0079] 其中，预设的特征排序方式包括但不限于：lightgbm算法、xgboost算法等，需要说明的是，树模型天然会对特征进行重要性排序，以分裂数据集，构建分支，进而根据分支的评分得到重要性排序，具体方式可根据实际需要进行选取，此处不做限制。

[0080] 需要说明的是，通过树模型计算出每个稳定性特征的排序后，将重要程度低于预设数值的稳定性指标剔除，以提高关键特征序列中特征的质量。

[0081] 优选地，本实施例采用xgboost算法对对稳定特征进行重要性排序筛选，具体过程如下：

[0082] 从树的根节点出发，每次选择一个稳定特征及其对应的特征值，并通过损失函数计算损失，将使得损失函数对应的损失最小稳定特征作为分裂节点，根据特征值对数据进行排序，然后按照稳定特征的特征值从小到大进行切分，比较采用每个分裂节点切分后的

目标函数大小，选择下降最大的分裂节点作为该稳定特征的最优切分点，最后比较分支最优切分点的目标函数下降值，选择下降最大的特征值作为最优切分点。依次执行上述过程，直到所有稳定特征参与拟合，将根节点到叶子节点之间稳定特征的顺序，作为稳定特征重要性排序。

[0083] S205：根据关键特征序列进行特征衍生，得到衍生特征。

[0084] 具体地，按照需要应用场景的类别和特征的复杂程度，从关键特征序列中选取若干稳定性特征，并对选取出的稳定性特征进行特征衍生，得到衍生特征。

[0085] 在本实施例中，特征衍生包括但不限于特征组合、特征交叉、图像特征生成、文本特征生成等。

[0086] 其中，特征组合具体可以通过特征两两之间的四则运算组合，逻辑与、或组合，多项式构造，特征自身与其均值作差等来实现。

[0087] 其中，特征交叉是指根据实际需求，对特征的属性进行交叉处理，得到多个新的特征，例如，对于，特征A有三个属性(A1，A2，A3)，特征B有两个属性(B1，B2)，采用特征A的属性对特征进行交叉，可以得到6个新的特征(A1，B2)、(A2，B2)、(A3，B2)、(B1，A1)、(B1，A2)和(B1，A3)。

[0088] 其中，图像特征生成是针对图像数据，采用图像元素组合或者图像迁移学习的方式，生成新的图像特征。

[0089] 本实施例中，通过获取基础数据，并按照预设标签类型对基础数据进行分类，得到初始类别信息，进而对每个初始类别信息进行缺失值处理，得到基础类别信息，再计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征，确保后续进行特征衍生的特征的质量，有利于提高后续特征衍生的效率，进而通过预设的特征排序方式，对稳定特征进行重要性排序筛选，得到关键特征序列，再根据关键特征序列进行特征衍生，得到衍生特征。实现从依据重要性进行排序的稳定特征序列中，快速进行特征衍生，提高了特征衍生的效率，以及得到的衍生特征的质量。

[0090] 在本实施例的一些可选的实现方式中，步骤S202中，对每个初始类别信息进行缺失值处理，得到基础类别信息包括：

[0091] 针对每个初始类别信息，获取初始类别信息中每个基础特征对应的特征值；

[0092] 对特征值进行数据校验，将未通过校验的特征值作为缺失值；

[0093] 对每个基础特征对应的缺失值进行统计，并将缺失值与所有特征值的比例超过预设比例的基础特征，作为无效特征，并从初始类别信息中移除无效特征，得到基础类别信息。

[0094] 其中，特征值是指基础特征对应的取值，具体可以是文本型、数值型和布尔型等，对特征值进行数据校验，具体包括但不限于：空值校验、数值规范性校验和数值唯一性校验。

[0095] 应理解，在缺失值与所有特征值的比例超过预设比例时，也即，缺失值较多时，认定该缺失值对应的基础特征存在质量问题，将该基础特征作为无效特征，从初始类别信息中移除，以避免后续该基础特征对模型训练产生负面影响。

[0096] 其中，空值检验可以正则表达式的方式实现，数值规范性校验通过将数值与预设规则进行匹配判断，数值唯一性校验是指判断是否存在相同的重复的数值。

[0097] 本实施例中，通过对一些质量不高的特征值进行剔除，确保得到基础类别信息中数据质量，减少后续稳定特征提取过程中的计算量，有利于提高后续提取稳定特征的效率。

[0098] 在本实施例的一些可选的实现方式中，步骤S203中，计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征包括：

[0099] 计算每个基础特征的信息值IV，并根据信息值IV进行特征筛选，得到关键特征；

[0100] 通过预设方式，计算关键特征的稳定度指标PSI，将稳定度指标PSI超过预设稳定度阈值的关键特征，作为稳定特征。

[0101] 具体地，每个基础特征包括至少一个属性特征，在本实施例中，基础特征为较复杂的数据，例如金融数据等，往往包含数千条属性特征，但是，针对具体的某项业务，大多数属性特征作用较小甚至没有关联，而过多的属性特征，在后续数据处理过程中，会导致耗时较长，效率极低，因而，需要对数据的属性特征进行筛选，以便提高后续处理的效率，同时，也避免不相关的数据对数据处理的准确度造成影响。本实施例中，采用信息值IV进行特征筛选，得到关键特征，进而计算关键特征的稳定度指标，将稳定度指标超过预设稳定度阈值的关键特征，确定为稳定特征。

[0102] 对数据属性特征进行的筛选方式，包括但不限于：皮尔逊相关系数(Pearson correlation coefficient)、基尼系数(Gini coefficient)、信息增益和信息值(Information Value，IV)等。作为一种优选方式，本实施例采用信息值IV进行特征筛选。

[0103] 其中，信息值IV是用于做特征选择计算评分卡时，表示每一个变量对目标变量来说有多少“信息”的量。

[0104] 本实施例采用信息值作为特征筛选的方式，计算基础特征中每个属性特征的信息值IV，并根据信息值IV筛选出关键特征的具体实现过程，可参考后续实施例的描述，为避免重复，此处不再赘述。

[0105] 值得说明的是，考虑到属性特征较多，为提高处理效率，可采用基尼系数，对属性特征进行降维，剔除对实际业务影响较小的属性特征，再通过计算信息值IV进行进一步筛选，有利于减少运算量，提高数据处理效率。

[0106] 本实施例中，根据信息值IV和稳定度指标PSI，对基础特征进行筛选，得到稳定特征，提高后续参与特征衍生的特征的质量，有利于提高特征衍生的效率，以及，有利于提高后续特征衍生得到的衍生特征的质量。

[0107] 在本实施例的一些可选的实现方式中，基础特征包括连续型的特征，计算每个基础特征的信息值IV包括：

[0108] 针对对基础特征中数据类型为连续型的特征，进行分箱处理，将连续型的特征转化为离散型特征；

[0109] 针对所有离散型特征进行独热编码，得到数字化变量；

[0110] 根据数字化变量，计算每个特征对应的信息值IV。

[0111] 具体地，每个基础特征包含多个属性特征，每个属性特征的类型，分为连续型和离散型两类，采用分箱法将连续性的属性特征离散化，进而对所有离散型的属性特征进行独热编码，并计算每个属性特征的信息值IV，以便后续根据信息值IV进行关键特征的筛选。

[0112] 其中，属性特征是基础特征中的具体一项特征，在金融领域，一个基础特征(数据)往往包含多个属性特征，例如，一条基础特征为用户信息数据，其包含用户姓名、用户性别、

联系方式和已办理业务等,每一项都为一个属性特征。

[0113] 其中,连续型的属性特征是指在一定区间内可以任意取值的属性特征,其数值是连续不断的,相邻两个数值可作无限分割,即可取无限个数值,例如,生产零件的规格尺寸,人体测量的身高、体重、胸围等为连续型的属性特征,其数值只能用测量或计量的方法取得。

[0114] 其中,离散型的属性特征是指特征值可以按一定顺序一一列举,通常以整数位取值的数据。如职工人数、工厂数、机器台数等,离散型属性特征的数值用计数的方法来获取。

[0115] 需要说明的是,本实施例中对于缺失数值的离散型属性特征进行空值填充,填充为特殊字符“NA”,避免属性特征无对应的特征值导致该条初始数据中后续训练过程中产生异常。

[0116] 进一步地,对于每一个基础特征,如果它有m个不同的属性特征,按照独热编码(one-hot编码)即得到m个二元特征。并且,这些特征值互斥,每次只有一个特征值被激活,被激活的特征值设置为1,其余不被激活的特征值则置为常数0,最终得到属性特征的每个特征值对应的基础数字编码。

[0117] 应理解,独热编码的方式能使原始状态的数据变成稀疏数据,能更好地解决数据挖掘对属性特征数据样本分类的问题,以及在一定程度上起到了扩充特征的作用,其中,原始状态的数据指初始数据及其属性特征的取值范围。

[0118] 例如,当属性特征为“性别”时,其特征值的取值范围包括“男”和“女”两个取值,即Gender=["male","female"],则性别为“男”对应的数字化编码为Gender=[1,0],性别为女对应的数字挂编码为Gender=[0,1]。

[0119] 值得说明的是,由于属性特征取值方式和取值范围不同,会影响后续特征筛选,而通过独热编码对不同属性特征的特征值采用统一编码方式,能使原始状态的特征值变成稀疏数据,避免了在特征筛选过程中由于不同特征值的取值方式不同对模型产生负面影响,从而有效提高筛选稳定指标的效果。

[0120] 在本实施例中,通过对属性特征进行预处理后,再进行独热编码,得到得到数字化变量,进而根据数字化变量计算信息值IV,减少了需要进行运算的数据量,有利于提高信息值IV计算的效率。

[0121] 在本实施例的一些可选的实现方式中,步骤S205之后,特征衍生方法还包括:

[0122] 对衍生特征进行稳定性校验;

[0123] 若衍生特征的稳定度超过预设稳定度阈值,保留衍生特征,并将衍生特征作为基础特征,否则,剔除衍生特征。

[0124] 具体地,在得到衍生特征后,需要进一步对衍生特征进行稳定性校验,将稳定度超过预设稳定度阈值的衍生特征,加入到基础特征中,在后续,可以继续参与到特征衍生之中,最大限度进行特征的衍生,提高衍生特征的丰富性,将稳定度未超过预设稳定度阈值的衍生特征进行剔除处理,以确保衍生特征的质量。

[0125] 应理解,对衍生特征进行稳定性校验与上述实施例中计算每个基础类别信息的基础特征的稳定度采用的方式相同,为避免重复,此处不再赘述。

[0126] 本实施例中,通过对衍生特征进行稳定性校验,确保衍生特征的质量,同时,将通过稳定性校验的衍生特征作为基础特征,进一步参与特征衍生,有利于提高特征衍生的效

率，同时，也使得特征更为丰富多样。

[0127] 应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不应对本发明实施例的实施过程构成任何限定。

[0128] 图3示出与上述实施例特征衍生方法一一对应的特征衍生装置的原理框图。如图3所示，该特征衍生装置包括信息获取模块31、数据处理模块32、特征筛选模块33、特征排序模块34和特征衍生模块35。各功能模块详细说明如下：

[0129] 信息获取模块31，用于获取基础数据，并按照预设标签类型对基础数据进行分类，得到初始类别信息；

[0130] 数据处理模块32，用于对每个初始类别信息进行缺失值处理，得到基础类别信息，其中，每个基础类别信息至少包括一个基础特征；

[0131] 特征筛选模块33，用于计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征；

[0132] 特征排序模块34，用于通过预设的特征排序方式，对稳定特征进行重要性排序筛选，得到关键特征序列；

[0133] 特征衍生模块35，用于根据关键特征序列进行特征衍生，得到衍生特征。

[0134] 可选地，数据处理模块32包括：

[0135] 特征值确定单元，用于针对每个初始类别信息，获取初始类别信息中每个基础特征对应的特征值；

[0136] 数据校验单元，用于对特征值进行数据校验，将未通过校验的特征值作为缺失值；

[0137] 基础类别信息确定单元，用于对每个基础特征对应的缺失值进行统计，并将缺失值与所有特征值的比例超过预设比例的基础特征，作为无效特征，并从初始类别信息中移除无效特征，得到基础类别信息。

[0138] 可选地，特征筛选模块33包括：

[0139] 信息值计算单元，用于计算每个基础特征的信息值IV，并根据信息值IV进行特征筛选，得到关键特征；

[0140] 稳定度计算单元，用于通过预设方式，计算关键特征的稳定度指标PSI，将稳定度指标PSI超过预设稳定度阈值的关键特征，作为稳定特征。

[0141] 可选地，基础特征包括连续型的特征，信息值计算单元包括：

[0142] 离散化子单元，用于针对对基础特征中数据类型为连续型的特征，进行分箱处理，将连续型的特征转化为离散型特征；

[0143] 数字化子单元，用于针对所有离散型特征进行独热编码，得到数字化变量；

[0144] 计算子单元，用于根据数字化变量，计算每个特征对应的信息值IV。

[0145] 可选地，特征衍生装置还包括：

[0146] 稳定度校验模块，用于对衍生特征进行稳定性校验；

[0147] 特征甑选模块，用于若衍生特征的稳定度超过预设稳定度阈值，保留衍生特征，并将衍生特征作为基础特征，否则，剔除衍生特征。

[0148] 关于特征衍生装置的具体限定可以参见上文中对于特征衍生方法的限定，在此不再赘述。上述特征衍生装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上

述各模块可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

[0149] 为解决上述技术问题，本申请实施例还提供计算机设备。具体请参阅图4，图4为本实施例计算机设备基本结构框图。

[0150] 所述计算机设备4包括通过系统总线相互通信连接存储器41、处理器42、网络接口43。需要指出的是，图中仅示出了具有组件连接存储器41、处理器42、网络接口43的计算机设备4，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。其中，本技术领域技术人员可以理解，这里的计算机设备是一种能够按照事先设定或存储的指令，自动进行数值计算和/或信息处理的设备，其硬件包括但不限于微处理器、专用集成电路(Application Specific Integrated Circuit，ASIC)、可编程门阵列(Field－Programmable Gate Array，FPGA)、数字处理器(Digital Signal Processor，DSP)、嵌入式设备等。

[0151] 所述计算机设备可以是桌上型计算机、笔记本、掌上电脑及云端服务器等计算设备。所述计算机设备可以与用户通过键盘、鼠标、遥控器、触摸板或声控设备等方式进行人机交互。

[0152] 所述存储器41至少包括一种类型的可读存储介质，所述可读存储介质包括闪存、硬盘、多媒体卡、卡型存储器(例如，SD或D界面显示存储器等)、随机访问存储器(RAM)、静态随机访问存储器(SRAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、可编程只读存储器(PROM)、磁性存储器、磁盘、光盘等。在一些实施例中，所述存储器41可以是所述计算机设备4的内部存储单元，例如该计算机设备4的硬盘或内存。在另一些实施例中，所述存储器41也可以是所述计算机设备4的外部存储设备，例如该计算机设备4上配备的插接式硬盘，智能存储卡(Smart Media Card，SMC)，安全数字(Secure Digital，SD)卡，闪存卡(Flash Card)等。当然，所述存储器41还可以既包括所述计算机设备4的内部存储单元也包括其外部存储设备。本实施例中，所述存储器41通常用于存储安装于所述计算机设备4的操作系统和各类应用软件，例如电子文件的控制的程序代码等。此外，所述存储器41还可以用于暂时地存储已经输出或者将要输出的各类数据。

[0153] 所述处理器42在一些实施例中可以是中央处理器(Central Processing Unit，CPU)、控制器、微控制器、微处理器、或其他数据处理芯片。该处理器42通常用于控制所述计算机设备4的总体操作。本实施例中，所述处理器42用于运行所述存储器41中存储的程序代码或者处理数据，例如运行电子文件的控制的程序代码。

[0154] 所述网络接口43可包括无线网络接口或有线网络接口，该网络接口43通常用于在所述计算机设备4与其他电子设备之间建立通信连接。

[0155] 本申请还提供了另一种实施方式，即提供一种计算机可读存储介质，所述计算机可读存储介质存储有界面显示程序，所述界面显示程序可被至少一个处理器执行，以使所述至少一个处理器执行如上述的特征衍生方法的步骤。

[0156] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质

(如ROM/RAM、磁碟、光盘)中,包括若干指令用以使得一台终端设备(可以是手机,计算机,服务器,空调器,或者网络设备等)执行本申请各个实施例所述的方法。

[0157] 显然,以上所描述的实施例仅仅是本申请一部分实施例,而不是全部的实施例,附图中给出了本申请的较佳实施例,但并不限制本申请的专利范围。本申请可以以许多不同的形式来实现,相反地,提供这些实施例的目的是使对本申请的公开内容的理解更加透彻全面。尽管参照前述实施例对本申请进行了详细的说明,对于本领域的技术人员来而言,其依然可以对前述各具体实施方式所记载的技术方案进行修改,或者对其中部分技术特征进行等效替换。凡是利用本申请说明书及附图内容所做的等效结构,直接或间接运用在其他相关的技术领域,均同理在本申请专利保护范围之内。

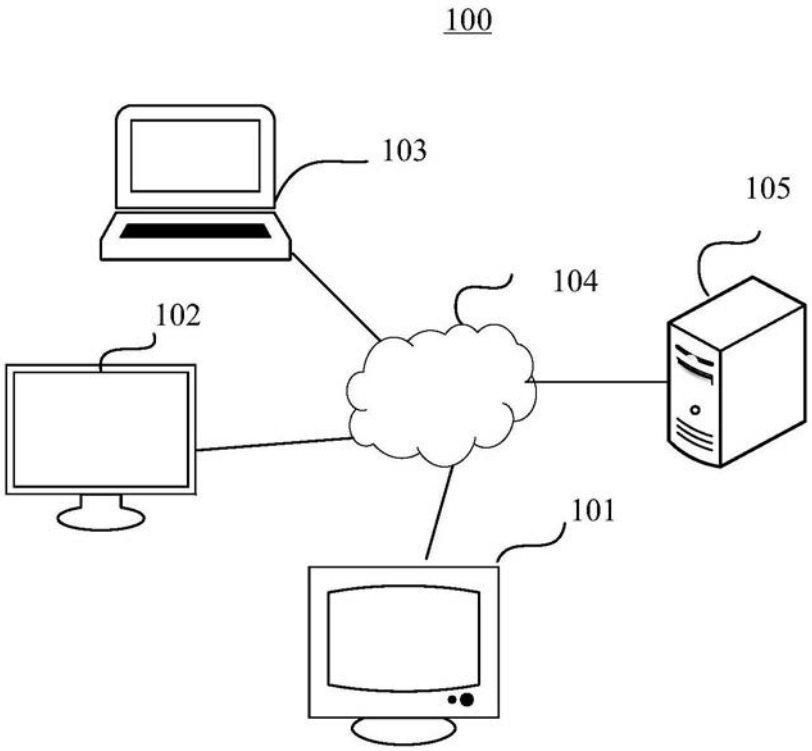


图1

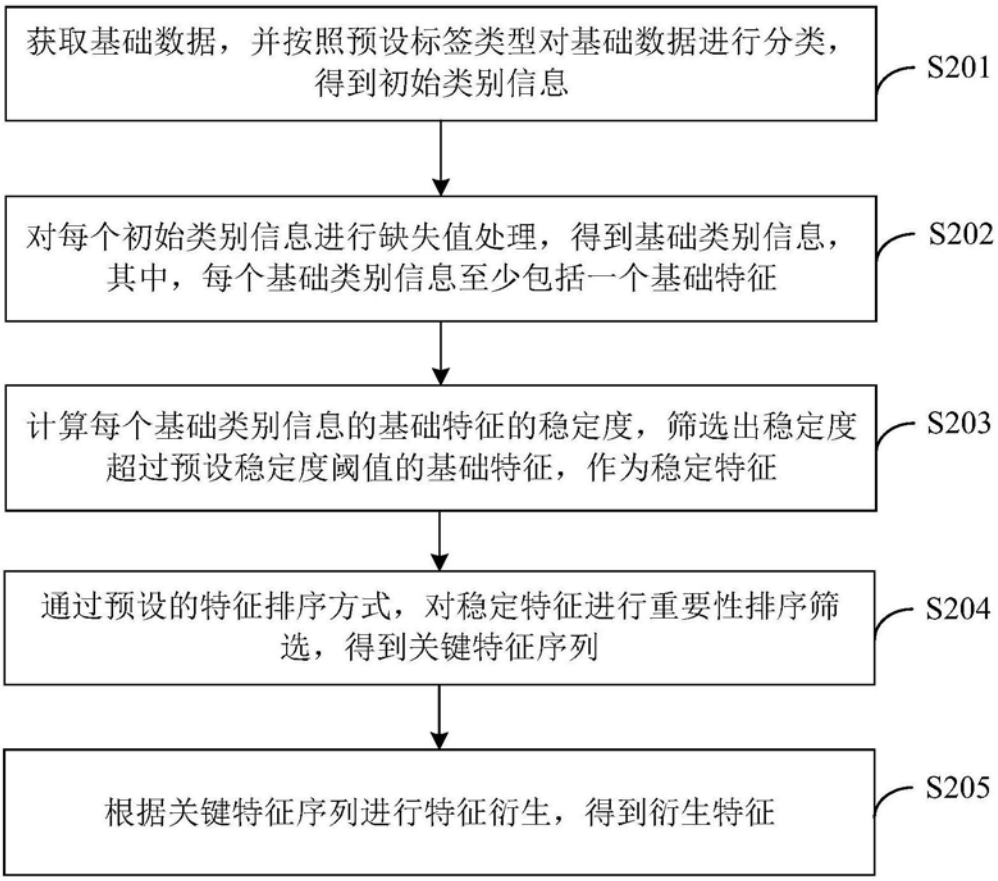

获取基础数据，并按照预设标签类型对基础数据进行分类，得到初始类别信息
S201
对每个初始类别信息进行缺失值处理，得到基础类别信息，其中，每个基础类别信息至少包括一个基础特征
S202
计算每个基础类别信息的基础特征的稳定度，筛选出稳定度超过预设稳定度阈值的基础特征，作为稳定特征
S203
通过预设的特征排序方式，对稳定特征进行重要性排序筛选，得到关键特征序列
S204
根据关键特征序列进行特征衍生，得到衍生特征
S205

图2

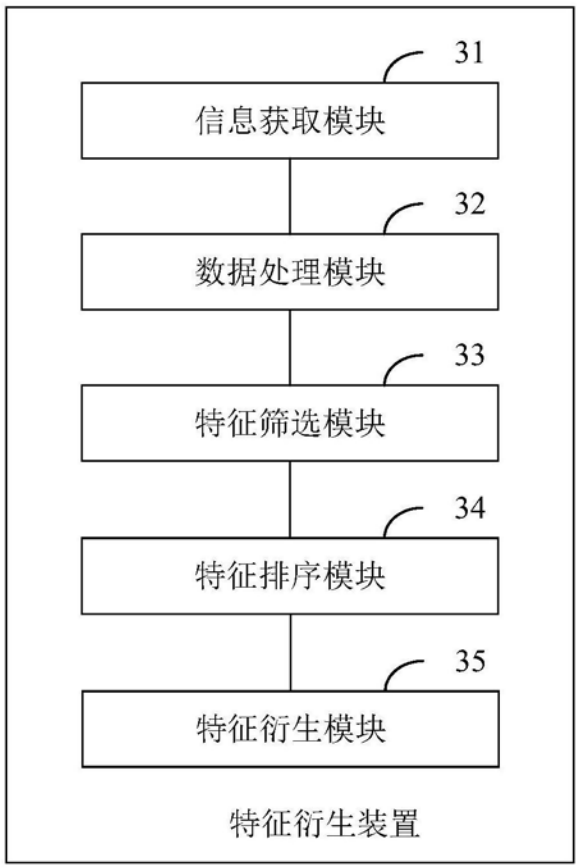

31
信息获取模块
32
数据处理模块
33
特征筛选模块
34
特征排序模块
35
特征衍生模块
特征衍生装置

图3

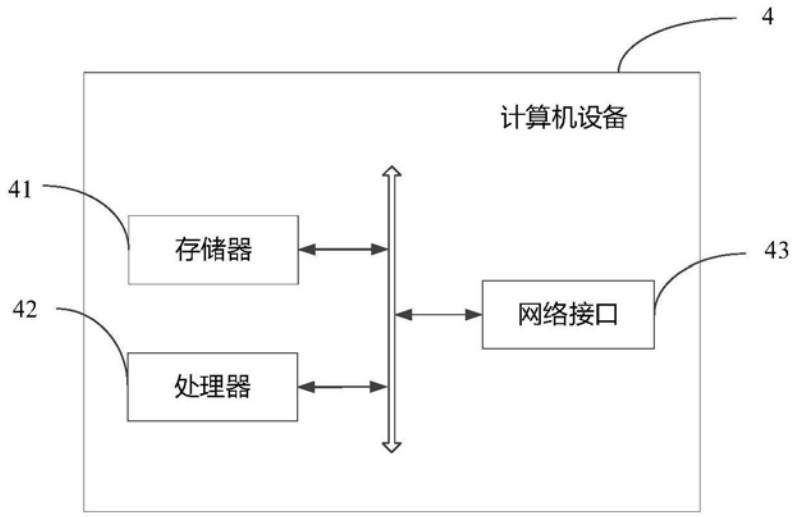

4
计算机设备
41
存储器
42
处理器
43
网络接口

图4