(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 108389125 B
(45)授权公告日 2022.06.07

(21)申请号 201810161147.7
(22)申请日 2018.02.27
(65)同一申请的已公布的文献号
申请公布号 CN 108389125 A
(43)申请公布日 2018.08.10
(73)专利权人 挖财网络技术有限公司
地址 310012 浙江省杭州市西湖区华星路96号第18层
(72)发明人 尤志强 潘琪 车曦
(74)专利代理机构 北京博思佳知识产权代理有限公司 11415
专利代理师 林祥
(51)Int.Cl.
G06Q 40/02(2012.01)

审查员 黎宾彬

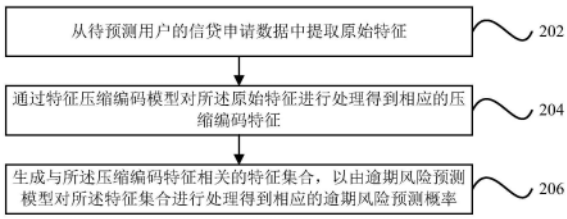
权利要求书3页 说明书11页 附图3页

(54)发明名称

信贷申请的逾期风险预测方法及装置

(57)摘要

本说明书一个或多个实施例提供一种信贷申请的逾期风险预测方法及装置，该方法可以包括：从待预测用户的信贷申请数据中提取原始特征；通过特征压缩编码模型对所述原始特征进行处理得到相应的压缩编码特征；其中，所述特征压缩编码模型由有标注样本数据对应的有标注样本特征和无标注样本数据对应的无标注样本特征进行无监督训练得到，所述有标注样本数据来源于已申请成功的信贷申请用户、所述无标注样本数据来源于已申请失败的信贷申请用户；生成与所述压缩编码特征相关的特征集合，以由逾期风险预测模型对所述特征集合进行处理得到相应的逾期风险预测概率。



CN 108389125 B

1.一种信贷申请的逾期风险预测方法,其特征在于,包括:

从待预测用户的信贷申请数据中提取原始特征;

通过特征压缩编码模型对所述原始特征进行处理得到相应的降维后的压缩编码特征;其中,所述特征压缩编码模型由有标注样本数据对应的有标注样本特征和无标注样本数据对应的无标注样本特征进行无监督训练得到,所述有标注样本数据来源于已申请成功的信贷申请用户、所述无标注样本数据来源于已申请失败的信贷申请用户;

生成与所述压缩编码特征相关的特征集合,以由逾期风险预测模型对所述特征集合进行处理得到相应的逾期风险预测概率;其中,所述有标注样本特征被所述特征压缩编码模型处理得到压缩编码样本特征,所述逾期风险预测模型由与所述压缩编码样本特征相关的样本特征集合、所述有标注样本特征对应的逾期状况标注信息进行有监督训练得到。

2.根据权利要求1所述的方法,其特征在于,当任一输入特征被输入所述特征压缩编码时,相应的输出特征包括所述特征压缩编码模型对所述输入特征进行压缩编码处理得到的隐变量。

3.根据权利要求2所述的方法,其特征在于,所述特征压缩编码模型包括:变分自编码器;所述隐变量由所述变分自编码器的编码层对所述输入特征进行压缩编码处理得到。

4.根据权利要求1所述的方法,其特征在于,

与所述压缩编码特征相关的特征集合包括:所述压缩编码特征;

与所述压缩编码样本特征相关的样本特征集合包括:所述压缩编码样本特征。

5.根据权利要求1所述的方法,其特征在于,

与所述压缩编码特征相关的特征集合包括:通过特征变换模型对所述压缩编码特征进行特征变换得到的变换后压缩编码特征;

与所述压缩编码样本特征相关的样本特征集合包括:通过所述特征变换模型对所述压缩编码样本特征进行特征变换得到的变换后压缩编码样本特征。

6.根据权利要求1所述的方法,其特征在于,所述特征集合还与所述原始特征相关,所述样本特征集合还与所述有标注样本特征相关。

7.根据权利要求6所述的方法,其特征在于,

与所述压缩编码特征相关的特征集合包括:所述原始特征和所述压缩编码特征;

与所述压缩编码样本特征相关的样本特征集合包括:所述有标注样本特征和所述压缩编码样本特征。

8.根据权利要求6所述的方法,其特征在于,

与所述压缩编码特征相关的特征集合包括:通过特征变换模型对所述原始特征和所述压缩编码特征构成的特征组合进行特征变换得到的变换后特征组合;

与所述压缩编码样本特征相关的样本特征集合包括:通过所述特征变换模型对所述有标注样本特征和所述压缩编码样本特征构成的特征组合进行特征变换得到的变换后样本特征组合。

9.根据权利要求5或8所述的方法,其特征在于,当任一输入特征被输入所述特征变换模型时,相应的输出特征包括:由所述特征变换模型对所述输入特征进行特征变换得到的具有区分性的特征和/或特征组合。

10.根据权利要求5或8所述的方法,其特征在于,所述特征变换模型包括:非线性特征

变换模型。

11.根据权利要求5或8所述的方法，其特征在于，所述特征变换模型包括：梯度提升决策树模型，所述梯度提升决策树模型通过迭代生成若干决策树；当任一输入特征被输入所述梯度提升决策树模型时，相应的输出特征由所述输入特征在所述决策树上落入的叶子节点而确定。

12.根据权利要求1所述的方法，其特征在于，所述逾期风险预测模型包括：线性分类器。

13.一种信贷申请的逾期风险预测装置，其特征在于，包括：

特征提取单元，从待预测用户的信贷申请数据中提取原始特征；

压缩编码单元，通过特征压缩编码模型对所述原始特征进行处理得到相应的降维后的压缩编码特征；其中，所述特征压缩编码模型由有标注样本数据对应的有标注样本特征和无标注样本数据对应的无标注样本特征进行无监督训练得到，所述有标注样本数据来源于已申请成功的信贷申请用户、所述无标注样本数据来源于已申请失败的信贷申请用户；

风险预测单元，生成与所述压缩编码特征相关的特征集合，以由逾期风险预测模型对所述特征集合进行处理得到相应的逾期风险预测概率；其中，所述有标注样本特征被所述特征压缩编码模型处理得到压缩编码样本特征，所述逾期风险预测模型由与所述压缩编码样本特征相关的样本特征集合、所述有标注样本特征对应的逾期状况标注信息进行有监督训练得到。

14.根据权利要求13所述的装置，其特征在于，当任一输入特征被输入所述特征压缩编码时，相应的输出特征包括所述特征压缩编码模型对所述输入特征进行压缩编码处理得到的隐变量。

15.根据权利要求14所述的装置，其特征在于，所述特征压缩编码模型包括：变分自编码器；所述隐变量由所述变分自编码器的编码层对所述输入特征进行压缩编码处理得到。

16.根据权利要求13所述的装置，其特征在于，

与所述压缩编码特征相关的特征集合包括：所述压缩编码特征；

与所述压缩编码样本特征相关的样本特征集合包括：所述压缩编码样本特征。

17.根据权利要求13所述的装置，其特征在于，

与所述压缩编码特征相关的特征集合包括：通过特征变换模型对所述压缩编码特征进行特征变换得到的变换后压缩编码特征；

与所述压缩编码样本特征相关的样本特征集合包括：通过所述特征变换模型对所述压缩编码样本特征进行特征变换得到的变换后压缩编码样本特征。

18.根据权利要求13所述的装置，其特征在于，所述特征集合还与所述原始特征相关，所述样本特征集合还与所述有标注样本特征相关。

19.根据权利要求18所述的装置，其特征在于，

与所述压缩编码特征相关的特征集合包括：所述原始特征和所述压缩编码特征；

与所述压缩编码样本特征相关的样本特征集合包括：所述有标注样本特征和所述压缩编码样本特征。

20.根据权利要求18所述的装置，其特征在于，

与所述压缩编码特征相关的特征集合包括：通过特征变换模型对所述原始特征和所述

压缩编码特征构成的特征组合进行特征变换得到的变换后特征组合；

与所述压缩编码样本特征相关的样本特征集合包括：通过所述特征变换模型对所述有标注样本特征和所述压缩编码样本特征构成的特征组合进行特征变换得到的变换后样本特征组合。

21.根据权利要求17或20所述的装置，其特征在于，当任一输入特征被输入所述特征变换模型时，相应的输出特征包括：由所述特征变换模型对所述输入特征进行特征变换得到的具有区分性的特征和/或特征组合。

22.根据权利要求17或20所述的装置，其特征在于，所述特征变换模型包括：非线性特征变换模型。

23.根据权利要求17或20所述的装置，其特征在于，所述特征变换模型包括：梯度提升决策树模型，所述梯度提升决策树模型通过迭代生成若干决策树；当任一输入特征被输入所述梯度提升决策树模型时，相应的输出特征由所述输入特征在所述决策树上落入的叶子节点而确定。

24.根据权利要求13所述的装置，其特征在于，所述逾期风险预测模型包括：线性分类器。

25.一种电子设备，其特征在于，包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器被配置为实现如权利要求1-12中任一项所述的方法。

信贷申请的逾期风险预测方法及装置

技术领域

[0001] 本说明书一个或多个实施例涉及数据处理技术领域，尤其涉及一种信贷申请的逾期风险预测方法及装置。

背景技术

[0002] 当用户提出信贷申请时，通过对该用户进行贷前的逾期风险预测，可以降低在完成借贷后发生逾期甚至形成坏账的概率。在相关技术中，可以通过设定判定规则，并基于该判定规则对用户进行逾期风险预测；但是，判定规则的建立需要耗费大量的时间和人力、物力成本，并且极度依赖于专家经验和主观因素，不仅效率极低，而且极容易受到环境因素的影响。

发明内容

[0003] 有鉴于此，本说明书一个或多个实施例提供一种信贷申请的逾期风险预测方法及装置。

[0004] 为实现上述目的，本说明书一个或多个实施例提供技术方案如下：

[0005] 根据本说明书一个或多个实施例的第一方面，提出了一种信贷申请的逾期风险预测方法，包括：

[0006] 从待预测用户的信贷申请数据中提取原始特征；

[0007] 通过特征压缩编码模型对所述原始特征进行处理得到相应的压缩编码特征；其中，所述特征压缩编码模型由有标注样本数据对应的有标注样本特征和无标注样本数据对应的无标注样本特征进行无监督训练得到，所述有标注样本数据来源于已申请成功的信贷申请用户、所述无标注样本数据来源于已申请失败的信贷申请用户；

[0008] 生成与所述压缩编码特征相关的特征集合，以由逾期风险预测模型对所述特征集合进行处理得到相应的逾期风险预测概率；其中，所述有标注样本特征被所述特征压缩编码模型处理得到压缩编码样本特征，所述逾期风险预测模型由与所述压缩编码样本特征相关的样本特征集合、所述有标注样本特征对应的逾期状况标注信息进行有监督训练得到。

[0009] 根据本说明书一个或多个实施例的第二方面，提出了一种信贷申请的逾期风险预测装置，包括：

[0010] 特征提取单元，从待预测用户的信贷申请数据中提取原始特征；

[0011] 压缩编码单元，通过特征压缩编码模型对所述原始特征进行处理得到相应的压缩编码特征；其中，所述特征压缩编码模型由有标注样本数据对应的有标注样本特征和无标注样本数据对应的无标注样本特征进行无监督训练得到，所述有标注样本数据来源于已申请成功的信贷申请用户、所述无标注样本数据来源于已申请失败的信贷申请用户；

[0012] 风险预测单元，生成与所述压缩编码特征相关的特征集合，以由逾期风险预测模型对所述特征集合进行处理得到相应的逾期风险预测概率；其中，所述有标注样本特征被所述特征压缩编码模型处理得到压缩编码样本特征，所述逾期风险预测模型由与所述压缩

编码样本特征相关的样本特征集合、所述有标注样本特征对应的逾期状况标注信息进行有监督训练得到。

[0013] 根据本说明书一个或多个实施例的第三方面，提出了一种电子设备，包括：

[0014] 处理器；

[0015] 用于存储处理器可执行指令的存储器；

[0016] 其中，所述处理器被配置为实现如上述实施例中任一项所述的方法。

附图说明

[0017] 图1是一示例性实施例提供的一种信贷申请的逾期风险预测系统的架构示意图。

[0018] 图2是一示例性实施例提供的一种信贷申请的逾期风险预测方法的流程图。

[0019] 图3是一示例性实施例提供的一种模型训练的示意图。

[0020] 图4是一示例性实施例提供的一种预测逾期风险发生概率的示意图。

[0021] 图5是一示例性实施例提供的一种设备的结构示意图。

[0022] 图6是一示例性实施例提供的一种信贷申请的逾期风险预测装置的框图。

具体实施方式

[0023] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本说明书一个或多个实施例相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本说明书一个或多个实施例的一些方面相一致的装置和方法的例子。

[0024] 需要说明的是：在其他实施例中并不一定按照本说明书示出和描述的顺序来执行相应方法的步骤。在一些其他实施例中，其方法所包括的步骤可以比本说明书所描述的更多或更少。此外，本说明书中所描述的单个步骤，在其他实施例中可能被分解为多个步骤进行描述；而本说明书中所描述的多个步骤，在其他实施例中也可能被合并为单个步骤进行描述。

[0025] 图1是一示例性实施例提供的一种信贷申请的逾期风险预测系统的架构示意图。如图1所示，该系统可以包括服务器11、网络12、若干电子设备，比如手机13、手机14和手机15等。

[0026] 服务器11可以为包含一独立主机的物理服务器，或者该服务器11可以为主机集群承载的虚拟服务器。手机13-15只是用户可以使用的一种类型的电子设备。实际上，用户显然还可以使用诸如下述类型的电子设备：平板设备、笔记本电脑、掌上电脑(PDAs，Personal Digital Assistants)、可穿戴设备(如智能眼镜、智能手表等)等，本说明书一个或多个实施例并不对此进行限制。

[0027] 而对于手机13-15与服务器11之间进行交互的网络12，可以包括多种类型的有线或无线网络。在一实施例中，该网络12可以包括公共交换电话网络(Public Switched Telephone Network，PSTN)和因特网。

[0028] 在运行过程中，服务器11可以运行信贷申请的逾期风险预测系统的服务器侧的程序，以实现对信贷申请的逾期风险预测功能。而电子设备可以运行信贷申请的逾期风险预

测系统的客户端侧的程序，可供实施与用户之间的人机交互操作，比如提交信贷申请数据、获知申请结果等；其中，该客户端侧的程序可以被预先安装在电子设备上，使得该客户端可以在该电子设备上被启动并运行；当然，当采用诸如HTML5技术的在线“客户端”时，无需在电子设备上安装相应的应用程序，即可获得并运行该客户端。当然，除了采用电子设备之外，用户也可以采用线下方式提交信贷申请数据、获知申请结果等，本说明书并不对此进行限制。

[0029] 下面结合实施例，对本申请的信贷申请的逾期风险预测方案进行说明。

[0030] 图2是一示例性实施例提供的一种信贷申请的逾期风险预测方法的流程图。如图2所示，该方法应用于服务器(比如图1所示的服务器11)，可以包括以下步骤：

[0031] 步骤202，从待预测用户的信贷申请数据中提取原始特征。

[0032] 在一实施例中，信贷申请数据可以由待预测用户主动提交，或者由第三方提供，本说明书并不对此进行限制。信贷申请数据可以包括与逾期风险预测相关的任意数据，比如该待预测用户的信息、所需申请的贷款的信息、用户历史行为信息等，本说明书并不对此进行限制。

[0033] 在一实施例中，在原始特征的抽取过程中，可以首先对待预测用户的信贷申请数据进行预处理，比如去除异常值、文本类型字段的数值化转换等，然后对预处理后的数据进行特征抽取、离散化处理、正则化处理等，最终得到上述的原始特征。当然，上述针对原始特征的抽取过程仅用于举例说明；实际上，从待预测用户的信贷申请数据中提取原始特征的过程，可以参照相关技术中的特征提取过程，本说明书并不对此进行限制。

[0034] 步骤204，通过特征压缩编码模型对所述原始特征进行处理得到相应的压缩编码特征；其中，所述特征压缩编码模型由有标注样本数据对应的有标注样本特征和无标注样本数据对应的无标注样本特征进行无监督训练得到，所述有标注样本数据来源于已申请成功的信贷申请用户、所述无标注样本数据来源于已申请失败的信贷申请用户。

[0035] 在一实施例中，通过将原始特征处理得到相应的压缩编码特征，在对该原始特征进行降维的同时，能够保留该原始特征所包含的本质特征信息，使得后续针对该压缩编码特征的处理过程中，既能够降低处理难度，又能够确保处理结果的准确程度。

[0036] 在一实施例中，当任一输入特征被输入所述特征压缩编码时，相应的输出特征包括所述特征压缩编码模型对所述输入特征进行压缩编码处理得到的隐变量。例如，当该任一输入特征为上述的原始特征时，输出特征可以为上述的压缩编码特征，那么该压缩编码特征可以为原始特征所包含的隐变量。

[0037] 在一实施例中，所述特征压缩编码模型可以包括：变分自编码器(Variational Auto-Encoder，简称VAE)，比如可以采用TensorFlow系统实现训练、支持增量学习；所述隐变量由所述变分自编码器的编码层对所述输入特征进行压缩编码处理得到。在完整的VAE模型中包括编码层(Encoder)和解码层(Decoder)；其中，编码层对n维的输入特征进行压缩编码处理后，形成m(m＜n)维的输出特征，在本说明书中即由编码层将n维的原始特征进行压缩编码处理得到m维的压缩编码特征，而无需应用解码层的解码处理。

[0038] 在其他实施例中，除了变分自编码器之外，还可以采用其他类型的模型对原始特征进行压缩编码处理，以得到上述的压缩编码特征，本说明书并不对此进行限制。

[0039] 在一实施例中，历史上已经提出过信贷申请的用户包括两类：已申请成功的信贷

申请用户和已申请失败的信贷申请用户;已申请成功的信贷申请用户可以形成相应的逾期状况标注数据,即该用户是否出现逾期、逾期时长、逾期金额等,而已申请失败的信贷申请用户则不存在相应的逾期状态标注数据。在一些情况下,已申请失败的信贷申请用户的数量甚至可能远大于已申请成功的信贷申请用户的数量,比如已申请成功的信贷申请用户在所有提出过信贷申请的用户中所占比例可能低于20%。可见,已申请失败的信贷申请用户对应的无标注样本数据,实际上包含了大量有意义的数据内容;例如,已申请失败的信贷申请用户中实际上存在很多实质上的优质用户(即不会发生逾期或发生逾期的概率较低)。因此,通过将有标注样本特征和无标注样本特征共同基于无监督训练方式形成特征压缩编码模型,相比于仅使用有标注样本特征及其逾期状况标注数据进行有监督训练明显具有更好的模型泛化能力,使得该特征压缩编码模型能够实现更加准确的特征压缩编码处理,从而提升对待预测用户的逾期风险预测准确度。

[0040] 步骤206,生成与所述压缩编码特征相关的特征集合,以由逾期风险预测模型对所述特征集合进行处理得到相应的逾期风险预测概率;其中,所述有标注样本特征被所述特征压缩编码模型处理得到压缩编码样本特征,所述逾期风险预测模型由与所述压缩编码样本特征相关的样本特征集合、所述有标注样本特征对应的逾期状况标注信息进行有监督训练得到。

[0041] 在一实施例中,与所述压缩编码特征相关的特征集合包括:所述压缩编码特征;与所述压缩编码样本特征相关的样本特征集合包括:所述压缩编码样本特征。换言之,当预先通过有标注样本特征对应的逾期状况标注信息和压缩编码样本特征训练得到逾期风险预测模型时,在针对待预测用户的逾期风险预测过程中,可以将上述的压缩编码特征作为该逾期风险预测模型的输入特征,以使得该逾期风险预测模型处理并输出相应的逾期风险预测概率,即该待预测用户出现逾期状况的概率。

[0042] 在一实施例中,与所述压缩编码特征相关的特征集合包括:通过特征变换模型对所述压缩编码特征进行特征变换得到的变换后压缩编码特征;与所述压缩编码样本特征相关的样本特征集合包括:通过所述特征变换模型对所述压缩编码样本特征进行特征变换得到的变换后压缩编码样本特征。换言之,当预先通过有标注样本特征对应的逾期状况标注信息和变换后压缩编码样本特征训练得到逾期风险预测模型时,在针对待预测用户的逾期风险预测过程中,可以将上述的变换后压缩编码特征作为该逾期风险预测模型的输入特征,以使得该逾期风险预测模型处理并输出相应的逾期风险预测概率,即该待预测用户出现逾期状况的概率。

[0043] 在一实施例中,所述特征集合还与所述原始特征相关,所述样本特征集合还与所述有标注样本特征相关;换言之,上述的特征集合可以同时与原始特征和压缩编码特征相关,而上述的样本特征集合可以同时与有标注样本特征和压缩编码样本特征相关。

[0044] 在一实施例中,与所述压缩编码特征相关的特征集合包括:所述原始特征和所述压缩编码特征;与所述压缩编码样本特征相关的样本特征集合包括:所述有标注样本特征和所述压缩编码样本特征。换言之,当预先通过有标注样本特征对应的逾期状况标注信息、有标注样本特征和压缩编码样本特征训练得到逾期风险预测模型时,在针对待预测用户的逾期风险预测过程中,可以将上述的原始特征和压缩编码特征作为该逾期风险预测模型的输入特征,以使得该逾期风险预测模型处理并输出相应的逾期风险预测概率,即该待预测

用户出现逾期状况的概率。虽然特征压缩编码模型在实施压缩编码处理的过程中，尽可能地保留了本质特征信息，但是仍然可能造成一定程度上的信息丢失，因而通过将有标注样本特征和压缩编码样本特征同时应用于训练逾期风险预测模型，可以弥补压缩编码处理可能造成的部分信息丢失的问题，从而既能够利用原始特征所包含内容的全面性，又能够利用压缩编码后低维特征具有更好的泛化能力的特性，有助于提升对逾期风险预测模型的训练效果。相应地，通过将原始特征和压缩编码特征同时输入逾期风险预测模型，同样能够充分发挥原始特征的内容全面性、压缩编码后低维特征具有更好的泛化能力的特性，有助于提升预测准确度。

[0045] 在一实施例中，与所述压缩编码特征相关的特征集合包括：通过特征变换模型对所述原始特征和所述压缩编码特征构成的特征组合进行特征变换得到的变换后特征组合；与所述压缩编码样本特征相关的样本特征集合包括：通过所述特征变换模型对所述有标注样本特征和所述压缩编码样本特征构成的特征组合进行特征变换得到的变换后样本特征组合。换言之，当预先通过有标注样本特征对应的逾期状况标注信息和变换后样本特征组合训练得到逾期风险预测模型时，在针对待预测用户的逾期风险预测过程中，可以将上述的变换后特征组合作为该逾期风险预测模型的输入特征，以使得该逾期风险预测模型处理并输出相应的逾期风险预测概率，即该待预测用户出现逾期状况的概率。

[0046] 在一实施例中，通过特征变换模型对压缩编码样本特征进行特征变换，可以发现多种有区分性的特征以及特征组合，保留信息量更大或与预测结果更相关的特征，省去人工寻找特征、特征组合的步骤，同时可使原始的连续的特征转换为离散特征。变换后压缩编码样本特征具有更佳的区分性，便于对不同类型的特征(如逾期风险相对更高的特征、逾期风险相对更低的特征等)予以更加准确的区分，从而使训练得到的逾期风险预测模型能够实现更加准确的逾期风险预测功能。类似地，通过特征变换模型对压缩编码特征进行特征变换，可以发现多种有区分性的特征以及特征组合，使得变换后压缩编码特征可以具有更佳的区分性，便于通过逾期风险预测模型实现准确的逾期风险预测。

[0047] 在一实施例中，当任一输入特征被输入所述特征变换模型时，相应的输出特征包括：由所述特征变换模型对所述输入特征进行特征变换得到的具有区分性的特征和/或特征组合。比如，当该输入特征为压缩编码样本特征时，特征变换模型可以对该压缩编码样本特征进行特征变换得到的具有区分性的特征和/或特征组合，以作为相应的变换后压缩编码样本特征。当该输入特征为压缩编码特征时，特征变换模型可以对该压缩编码特征进行特征变换得到的具有区分性的特征和/或特征组合，以作为相应的变换后压缩编码特征。

[0048] 在一实施例中，所述特征变换模型可以包括：非线性特征变换模型。例如，该特征变换模型可以包括：梯度提升决策树(Gradient Boosting Decision Tree，简称GBDT)模型，所述梯度提升决策树模型通过迭代生成若干决策树，具体可以在每次迭代时都在减少残差的梯度方向生成一棵决策树，每棵树的每个叶子节点对应变换后的一维特征；因此，当任一输入特征被输入所述梯度提升决策树模型时，可以根据输入特征在所述决策树上落入的叶子节点确定出相应的输出特征。GBDT模型本身可以应用于回归或分类，但本说明书中利用了GBDT模型能够生成具有区分性的特征或特征组合的特点，将其应用于特征变换操作。除了GBDT模型之外，还可以采用其他类型的非线性特征变换模型，本说明书并不对此进行限制；例如，可以通过DNN(Deep Neural Network，深度神经网络)模型实现上述的非线性

特征变换。

[0049] 在一实施例中，所述逾期风险预测模型可以包括：线性分类器；例如，该线性分类器可以基于逻辑回归(Logistic Regression)模型训练得到。当然，还可以采用其他模型训练，比如基于线性模型加交叉项的因子分解机(Factorization Machine，简称FM)模型训练得到。本说明书并不对此进行限制；在一些情况下，甚至可能采用线性分类器之外的其他分类器，本说明书并不对此进行限制。

[0050] 在一实施例中，可以将特征压缩编码模型与逾期风险预测模型进行集成应用，或者将特征压缩编码模型、特征变换模型与逾期风险预测模型进行集成应用。其中，特征压缩编码模型采用无监督方式进行训练，与采用有监督方式训练得到的特征变换模型、逾期风险预测模型进行集成，可以整体上形成类似于半监督形式的stacking集成算法，可以同时发挥多个模型的优势和特点、取长补短，从而达到较之单个模型或算法更佳的处理效果。

[0051] 为了便于理解，下面以金融机构在信贷申请过程中实施的逾期风险预测操作为例，对本说明书一个或多个实施例的技术方案进行说明。假定如图1所示的服务器11上配置有信贷申请的逾期风险预测系统的服务端，而用户X使用的手机13上配置有信贷申请的逾期风险预测系统的客户端，使得用户X可以基于该客户端发起信贷申请，而服务端可以针对该信贷申请实施相应的逾期风险预测操作，以预测该用户X发生逾期的概率，从而据此确定是否通过或拒绝该用户X发起的信贷申请。

[0052] 按照上述逾期风险预测操作的发生顺序，可以将整个过程划分为两个阶段：第一阶段为模型训练阶段，第二阶段为风险预测阶段；下面针对这两个阶段分别进行详细描述。

[0053] 图3是一示例性实施例提供的一种模型训练的示意图。如图3所示，该模型训练的过程发生于服务器11上运行的服务端，可以包括以下步骤：

[0054] 步骤①，根据获取到的全量样本数据，形成相应的样本特征。

[0055] 在一实施例中，全量样本数据可以包括全量历史数据的至少一部分，可以根据实际情况进行选择。例如，可以选取最近3个月的历史数据、以作为该全量样本数据。通过设定一定数值的时间窗口，并在时间轴上移动该时间窗口，可以将对应于该时间窗口的时间段产生的历史数据作为上述的全量样本数据，比如该时间窗口可以为上述的3个月或者其他任意时长。同时，通过定期移动该时间窗口，可以对全量样本数据进行更新，使得相应更新训练本说明书涉及到的至少一个模型，以适应于实际情况的变化。

[0056] 在一实施例中，全量样本数据中的“全量”是针对样本数据的类型而言；具体地，全量样本数据可以包括两种类型的样本数据：有标注样本数据和无标注样本数据，还包括该有标注样本数据对应的标注信息。其中，有标注样本数据是指提出过信贷申请且申请成功的用户对应的信贷申请数据，而标注信息是指这些申请成功的用户对应的逾期状况标注信息，比如该逾期状况标注信息可以包括未发生逾期、发生过逾期、逾期时长、逾期金额、逾期次数等；无标注样本数据是指提出过信贷申请但申请失败的用户对应的信贷申请数据，由于申请失败因而并未能够成功放款，因而不存在相应的标注信息。

[0057] 实际上，在已经申请成功的用户中，必然存在出现逾期的违约用户，而在申请失败的用户中，也必然存在并不会或极低概率出现逾期的优质用户，这些都意味着对于相关用户的逾期风险预测尚不到位。因此，通过采用上述的全量样本数据作为模型训练样本，可以兼顾“成功识别出优质用户并放款”、“成功识别出非优质用户并拒绝放款”、“未识别出优质

用户并拒绝放款”、“未识别出非优质用户并放款”等多种情况，有助于提升模型训练的全面性和准确性。

[0058] 在一实施例中，通过对有标注样本数据和无标注样本数据实施相关处理，可以提取出相应的样本特征，即有标注样本数据对应的有标注样本特征A、无标注样本数据对应的无标注样本特征。例如，上述的相关处理可以包括对有标注样本数据、无标注样本数据分别实施预处理，比如去除异常值、文本类型字段的数值化转换；以及，上述的相关处理可以包括对预处理后的有标注样本数据、无标注样本数据分别实施特征抽取，还包括特征的离散化处理、正则化等操作，使得最终得到的有标注样本特征A、无标注样本特征为数值化的字段特征。

[0059] 步骤②，根据样本特征训练VAE模型。

[0060] 在一实施例中，根据步骤①得到的有标注样本特征A和无标注样本特征，可以共同用于训练VAE模型。由于VAE模型的训练方式为无监督训练，使得无标注样本特征能够被应用于对该VAE模型的训练，并最终应用于对逾期风险的预测操作中，以提升对逾期风险的预测准确率。

[0061] 步骤③，通过训练得到的VAE模型对有标注样本特征A进行处理，得到压缩编码样本特征A’，以构成样本特征组合A+A’。

[0062] 在一实施例中，VAE模型包括两个部分：编码层和解码层；其中，编码层用于对输入特征(比如上述的有标注样本特征A)进行压缩编码处理，得到相应的压缩编码样本特征A’，该压缩编码样本特征A’用于表达有标注样本特征A包含的隐变量，使得该压缩编码样本特征A’的维度低于有标注样本特征A的情况下，可以通过该隐变量保留了有标注样本特征A的关键性、具有决定性影响的信息。

[0063] 在相关技术中，解码层用于对压缩编码样本特征A’进行还原，以得到有标注样本特征A或其近似特征；而在本说明书的技术方案中，主要应用编码层对有标注样本特征A的压缩编码处理，而不需要对解码层进行应用。

[0064] 步骤④，根据样本特征组合A+A’、有标注样本数据对应的标注信息，训练特征变换模型G。

[0065] 在一实施例中，对特征变换模型G的训练可以为有监督训练，因而需要应用有标注样本数据对应的标注信息。特征变换模型G的训练样本可以为上述的样本特征组合A+A’，即同时将有标注样本特征A与压缩编码样本特征A’应用于对特征变换模型G的训练过程；其中，虽然压缩编码样本特征A’保留了有标注样本特征A的关键信息，但是仍然可能丢失至少一部分有用信息，因而通过采用样本特征组合A+A’，可使该至少一部分有用信息得以参与至对于特征变换模型G的训练过程，从而有助于提升特征变换模型G的准确性。

[0066] 在一实施例中，特征变换模型G可以采用GBDT模型。GBDT模型基于boosting机制，可以在每次迭代时都在减少残差的梯度方向新创建一棵决策树，每棵树的每个叶子节点对应一维特征，因而基于这些决策树可以获得具有区分性的特征和/或特征组合，省去了人工寻找特征、特征组合的步骤。

[0067] 在一实施例中，当决策树的数量小于预设数值时，这些决策树上发生的树分裂主要体现了对于多数样本具有区分度的特征；而此后继续生成的决策树，其树分裂主要体现的是对于经过先前的决策树后残差仍然较大的少数样本具有区分度的特征。在本说明书的

技术方案中，可以优先选用在整体上具有区分度的特征，在此基础上可以选择性地采用针对少数样本具有区分度的特征。

[0068] 在一实施例中，特征变换模型G可以采用其他类型的模型，比如DNN等非线性特征变换模型，本说明书并不对此进行限制。

[0069] 步骤⑤，通过训练得到的特征变换模型G对样本特征组合A+A'进行处理，得到变换后样本特征组合。

[0070] 在一实施例中，通过特征变换模型G对样本特征组合A+A'进行处理，可以得到具有区分性的离散化特征，即变换后样本特征组合，以便于后续对于线性分类器C的高效、可靠训练。

[0071] 在一实施例中，可以分别将样本特征组合A+A'中的各个特征输入训练好的GBDT模型中，并根据这些特征在各个决策树上落入的叶子节点，实现相应的特征变换，从而将样本特征组合A+A'处理为相应的变换后样本特征组合。

[0072] 步骤⑥，根据变换后样本特征组合、有标注样本数据对应的标注信息，训练线性分类器C。

[0073] 在一实施例中，对线性分类器C的训练可以为有监督训练，因而需要应用有标注样本数据对应的标注信息。

[0074] 在一实施例中，线性分类器C可以基于逻辑回归(Logistic Regression)模型训练得到。当然，还可以采用其他模型训练得到上述的线性分类器，本说明书并不对此进行限制；在一些情况下，甚至可能采用线性分类器之外的其他分类器，本说明书并不对此进行限制。

[0075] 通过如图3所示的上述步骤，可以基于无监督训练得到VAE模型、基于有监督训练得到特征变换模型G、基于有监督训练得到线性分类器C，从而整体上组成了半监督模式的stacking集成算法。此外，还可以采用与全量样本数据相区别的验证样本数据对上述半监督模式的stacking集成算法进行验证，以及采用与全量样本数据、验证样本数据相区别的测试样本数据对上述半监督模式的stacking集成算法进行测试，以确保该半监督模式的stacking集成算法符合应用需求。而实际上，上述半监督模式的stacking集成算法可以混合多种模型的优势、取长补短，其效果远优于采用单个模型或者加法模型、投票模型等其他形式的集成算法，能够很好的满足实际应用需求，基于信贷申请数据准确地预测出用户的逾期风险。

[0076] 例如，基于图3训练得到的VAE模型、特征变换模型G和线性分类器C，可以对用户X提交的信贷申请数据进行处理，以确定出用户X的逾期风险发生概率。相应地，图4是一示例性实施例提供的一种预测逾期风险发生概率的示意图；如图4所示，该预测过程发生于服务器11上运行的服务端，可以包括以下步骤：

[0077] 步骤(1)，根据用户X对应的信贷申请数据，形成相应的原始特征B。

[0078] 在一实施例中，与图3所示实施例中的步骤①相类似的，通过对用户X的信贷申请数据进行数据预处理、特征抽取等操作，可以获得相应的原始特征B。

[0079] 在一实施例中，本实施例中的用户X，以及本说明书中其他实施例中提及的用户，均可以为提出信贷申请的任意个人或企业机构等，本说明书并不对此进行限制。

[0080] 步骤(2)，通过训练得到的VAE模型对原始特征B进行处理，得到压缩编码特征B'，

以构成特征组合B+B’。

[0081] 在一实施例中，通过训练得到的VAE模型对原始特征B进行处理，可以得到相应的压缩编码特征B’，该压缩编码特征B’可以在降维的同时、保留原始特征B的关键信息。

[0082] 步骤(3)，通过训练得到的训练特征变换模型G对特征组合B+B’进行处理，得到变换后特征组合。

[0083] 在一实施例中，通过采用特征组合B+B’，既可以发挥压缩编码特征B’在低维度下具有更好泛化能力的特性，又能够发挥原始特征B所包含信息的全面性，便于对用户X实现更为准确的逾期风险预测。

[0084] 在一实施例中，当特征变换模型G为GBDT模型时，可以将原始特征B输入该GBDT模型迭代生成的决策树中，并根据该原始特征B落入的叶子节点，得到相应的压缩编码特征B’。

[0085] 步骤(4)，通过训练得到的线性分类器C对变换后特征组合进行处理，得到针对该用户X的逾期风险预测概率。

[0086] 在一实施例中，通过将变换后特征组合输入训练得到的线性分类器C，可由线性分类器C进行处理得到相应的输出数据，即针对用户X的逾期风险预测概率。

[0087] 在一实施例中，根据预先定义的概率阈值，当用户X的逾期风险预测概率大于该概率阈值时，可以判定用户X很可能发生逾期，因而可以拒绝用户X的信贷申请；而当用户X的逾期风险预测概率不大于该概率阈值时，可以判定用户X可能并不会发生逾期，因而可以确认用户X的信贷申请通过审批。

[0088] 图5是一示例性实施例提供的一种设备的示意结构图。请参考图5，在硬件层面，该设备包括处理器502、内部总线504、网络接口506、内存508以及非易失性存储器510，当然还可能包括其他业务所需要的硬件。处理器502从非易失性存储器510中读取对应的计算机程序到内存508中然后运行，在逻辑层面上形成信贷申请的逾期风险预测装置。当然，除了软件实现方式之外，本说明书一个或多个实施例并不排除其他实现方式，比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限定于各个逻辑单元，也可以是硬件或逻辑器件。

[0089] 请参考图6，在软件实施方式中，该信贷申请的逾期风险预测装置可以包括：

[0090] 特征提取单元601，从待预测用户的信贷申请数据中提取原始特征；

[0091] 压缩编码单元602，通过特征压缩编码模型对所述原始特征进行处理得到相应的压缩编码特征；其中，所述特征压缩编码模型由有标注样本数据对应的有标注样本特征和无标注样本数据对应的无标注样本特征进行无监督训练得到，所述有标注样本数据来源于已申请成功的信贷申请用户、所述无标注样本数据来源于已申请失败的信贷申请用户；

[0092] 风险预测单元603，生成与所述压缩编码特征相关的特征集合，以由逾期风险预测模型对所述特征集合进行处理得到相应的逾期风险预测概率；其中，所述有标注样本特征被所述特征压缩编码模型处理得到压缩编码样本特征，所述逾期风险预测模型由与所述压缩编码样本特征相关的样本特征集合、所述有标注样本特征对应的逾期状况标注信息进行有监督训练得到。

[0093] 可选的，当任一输入特征被输入所述特征压缩编码时，相应的输出特征包括所述特征压缩编码模型对所述输入特征进行压缩编码处理得到的隐变量。

[0094] 可选的，所述特征压缩编码模型包括：变分自编码器；所述隐变量由所述变分自编码器的编码层对所述输入特征进行压缩编码处理得到。

[0095] 可选的，

[0096] 与所述压缩编码特征相关的特征集合包括：所述压缩编码特征；

[0097] 与所述压缩编码样本特征相关的样本特征集合包括：所述压缩编码样本特征。

[0098] 可选的，

[0099] 与所述压缩编码特征相关的特征集合包括：通过特征变换模型对所述压缩编码特征进行特征变换得到的变换后压缩编码特征；

[0100] 与所述压缩编码样本特征相关的样本特征集合包括：通过所述特征变换模型对所述压缩编码样本特征进行特征变换得到的变换后压缩编码样本特征。

[0101] 可选的，所述特征集合还与所述原始特征相关，所述样本特征集合还与所述有标注样本特征相关。

[0102] 可选的，

[0103] 与所述压缩编码特征相关的特征集合包括：所述原始特征和所述压缩编码特征；

[0104] 与所述压缩编码样本特征相关的样本特征集合包括：所述有标注样本特征和所述压缩编码样本特征。

[0105] 可选的，

[0106] 与所述压缩编码特征相关的特征集合包括：通过特征变换模型对所述原始特征和所述压缩编码特征构成的特征组合进行特征变换得到的变换后特征组合；

[0107] 与所述压缩编码样本特征相关的样本特征集合包括：通过所述特征变换模型对所述有标注样本特征和所述压缩编码样本特征构成的特征组合进行特征变换得到的变换后样本特征组合。

[0108] 可选的，当任一输入特征被输入所述特征变换模型时，相应的输出特征包括：由所述特征变换模型对所述输入特征进行特征变换得到的具有区分性的特征和/或特征组合。

[0109] 可选的，所述特征变换模型包括：非线性特征变换模型。

[0110] 可选的，所述特征变换模型包括：梯度提升决策树模型，所述梯度提升决策树模型通过迭代生成若干决策树；当任一输入特征被输入所述梯度提升决策树模型时，相应的输出特征由所述输入特征在所述决策树上落入的叶子节点而确定。

[0111] 可选的，所述逾期风险预测模型包括：线性分类器。

[0112] 上述实施例阐明的系统、装置、模块或单元，具体可以由计算机芯片或实体实现，或者由具有某种功能的产品来实现。一种典型的实现设备为计算机，计算机的具体形式可以是个人计算机、膝上型计算机、蜂窝电话、相机电话、智能电话、个人数字助理、媒体播放器、导航设备、电子邮件收发设备、游戏控制台、平板计算机、可穿戴设备或者这些设备中的任意几种设备的组合。

[0113] 在一个典型的配置中，计算机包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[0114] 内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。内存是计算机可读介质的示例。

[0115] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带、磁盘存储、量子存储器、基于石墨烯的存储介质或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体(transitory media)，如调制的数据信号和载波。

[0116] 还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

[0117] 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

[0118] 在本说明书一个或多个实施例使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0119] 应当理解，尽管在本说明书一个或多个实施例可能采用术语第一、第二、第三等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本说明书一个或多个实施例范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[0120] 以上所述仅为本说明书一个或多个实施例的较佳实施例而已，并不用以限制本说明书一个或多个实施例，凡在本说明书一个或多个实施例的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本说明书一个或多个实施例保护的范围之内。

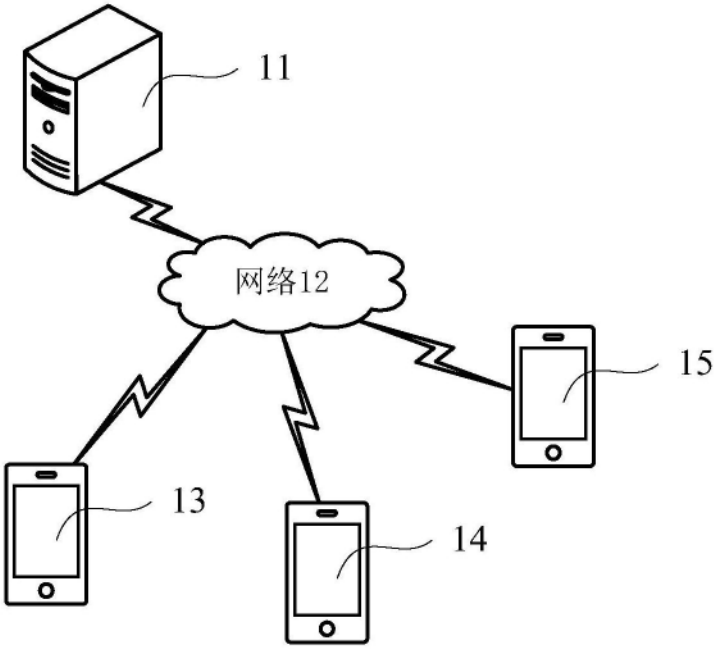

11
网络12
15
13
14

图1

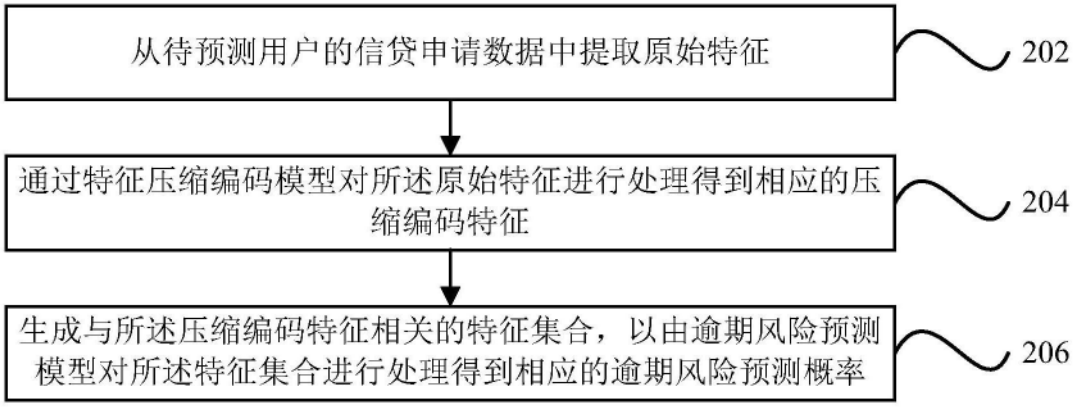

从待预测用户的信贷申请数据中提取原始特征
202
通过特征压缩编码模型对所述原始特征进行处理得到相应的压缩编码特征
204
生成与所述压缩编码特征相关的特征集合，以由逾期风险预测模型对所述特征集合进行处理得到相应的逾期风险预测概率
206

图2

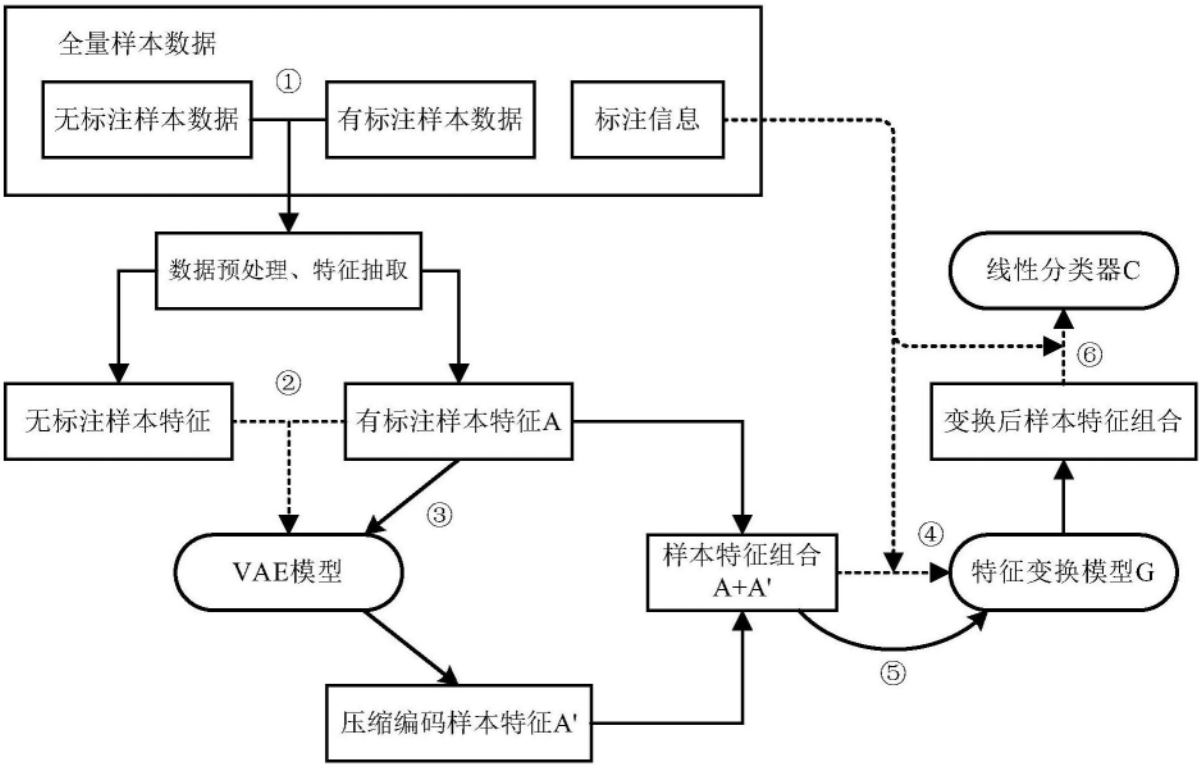

全量样本数据
①
无标注样本数据
有标注样本数据
标注信息
数据预处理、特征抽取
线性分类器C
⑥
②
无标注样本特征
有标注样本特征A
变换后样本特征组合
③
④
VAE模型
样本特征组合
A+A'
特征变换模型G
⑤
压缩编码样本特征A'

图3

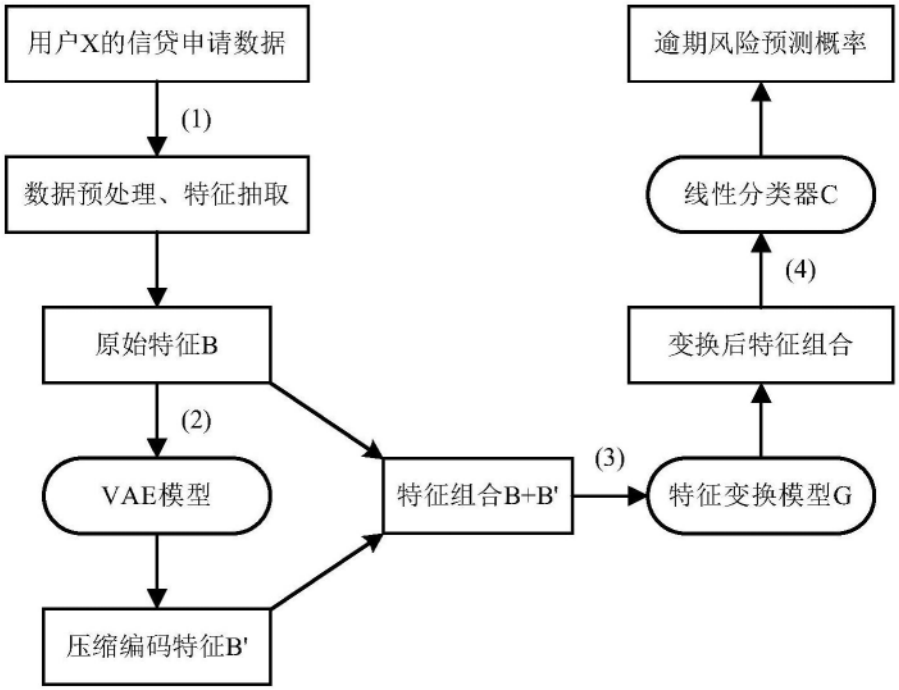

用户X的信贷申请数据
逾期风险预测概率
(1)
数据预处理、特征抽取
线性分类器C
(4)
原始特征B
变换后特征组合
(2)
(3)
VAE模型
特征组合B+B'
特征变换模型G
压缩编码特征B'

图4

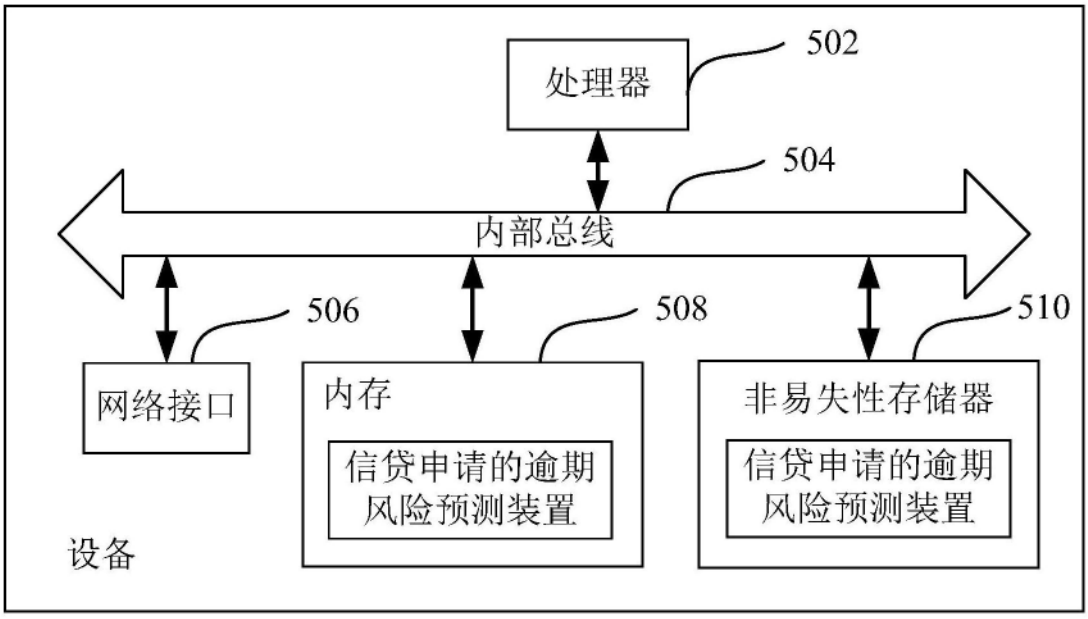

处理器
502
504
内部总线
506
508
510
网络接口
内存
信贷申请的逾期风险预测装置
非易失性存储器
信贷申请的逾期风险预测装置
设备

图5

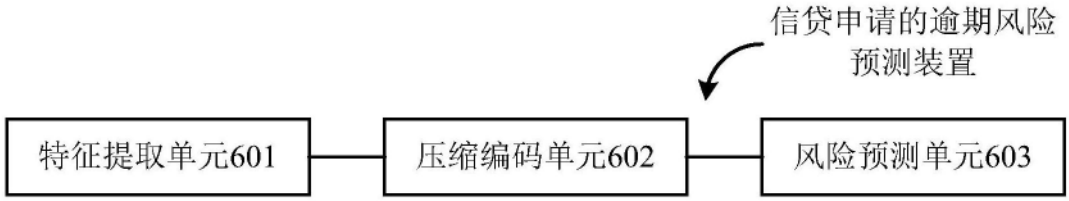

信贷申请的逾期风险预测装置
特征提取单元601
压缩编码单元602
风险预测单元603

图6