(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 109102023 A
(43)申请公布日 2018.12.28

(21)申请号 201810924268.2

(22)申请日 2018.08.14

(71)申请人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 郑毅 张鹏 潘健民

(74)专利代理机构 北京智信禾专利代理有限公司 11637
代理人 吴肖肖

(51)Int.Cl.
G06K 9/62(2006.01)

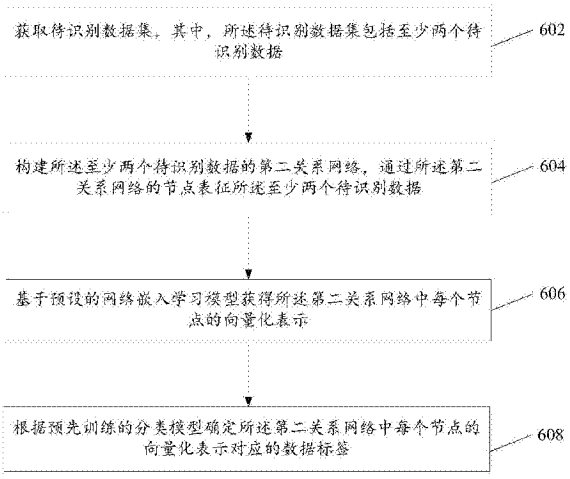
权利要求书3页 说明书10页 附图8页

(54)发明名称

一种分类模型生成方法及装置、一种数据识别方法及装置

(57)摘要

本申请提供的一种分类模型生成方法及装置、一种数据识别方法及装置，其中，所述数据识别方法包括获取待识别数据集；构建所述至少两个待识别数据的第二关系网络，通过所述第二关系网络的节点表征所述至少两个待识别数据；基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示；根据预先训练的分类模型确定所述第二关系网络中每个节点的向量化表示对应的数据标签。



CN 109102023 A

1.一种分类模型生成方法，其特征在于，包括：

获取训练样本数据集，其中，所述训练样本数据集中包括至少两个样本数据以及每个所述样本数据对应的样本标签；

构建所述至少两个样本数据的第一关系网络，通过所述第一关系网络的节点表征所述至少两个样本数据；

基于预设的网络嵌入学习模型获得所述第一关系网络中每个节点的向量化表示；

通过所述训练样本数据集对分类模型进行训练，得到所述分类模型，所述分类模型使得所述样本标签与所述第一关系网络中每个节点的向量化表示相关联。

2.根据权利要求1所述的方法，其特征在于，获取训练样本数据集包括：

按照预设时间间隔获取训练样本数据集。

3.根据权利要求1所述的方法，其特征在于，基于预设的网络嵌入学习模型获得所述第一关系网络中每个节点的向量化表示包括：

采用随机游走算法对所述第一关系网络中每个节点进行序列采样，并生成第一节点序列；

基于预设的网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示。

4.根据权利要求3所述的方法，其特征在于，基于预设的网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示包括：

基于Node2vec网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示。

5.根据权利要求3所述的方法，其特征在于，基于预设的网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示包括：

基于DeepWalk网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示。

6.根据权利要求4所述的方法，其特征在于，基于Node2vec网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示包括：

所述Node2vec网络嵌入学习模型基于Word2vec的SkipGram框架将所述节点序列中的每个节点进行向量化表示。

7.根据权利要求1所述的方法，其特征在于，所述分类模型包括XGBoost模型、随机森林模型、支持向量机模型或逻辑回归模型。

8.一种数据识别方法，其特征在于，包括：

获取待识别数据集，其中，所述待识别数据集包括至少两个待识别数据；

构建所述至少两个待识别数据的第二关系网络，通过所述第二关系网络的节点表征所述至少两个待识别数据；

基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示；

根据预先训练的分类模型确定所述第二关系网络中每个节点的向量化表示对应的数据标签。

9.根据权利要求8所述的方法，其特征在于，基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示包括：

采用随机游走算法对所述第二关系网络中每个节点进行序列采样，并生成第二节点序列；

根据预设的网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示。

10.根据权利要求9所述的方法，其特征在于，根据预设的网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示包括：

基于Node2vec网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示。

11.根据权利要求9所述的方法，其特征在于，根据预设的网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示包括：

基于DeepWalk网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示。

12.根据权利要求10所述的方法，其特征在于，基于Node2vec网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示包括：

所述Node2vec网络嵌入学习模型基于Word2vec的SkipGram框架将所述第二节点序列中的每个节点进行向量化表示。

13.根据权利要求8所述的方法，其特征在于，所述数据标签包括白数据标签和黑数据标签。

14.根据权利要求13所述的方法，其特征在于，还包括：

若所述数据标签为黑数据标签，则对所述数据标签对应的所述第二关系网络中节点的向量化表示进行后续处理。

15.根据权利要求8所述的方法，其特征在于，所述待识别数据集包括通过所述分类模型判断是否存在洗钱行为的待识别账户的集合。

16.一种分类模型生成装置，其特征在于，包括：

第一获取模块，被配置为获取训练样本数据集，其中，所述训练样本数据集中包括至少两个样本数据以及每个所述样本数据对应的样本标签；

第一构建模块，被配置为构建所述至少两个样本数据的第一关系网络，通过所述第一关系网络的节点表征所述至少两个样本数据；

第一学习模块，被配置为基于预设的网络嵌入学习模型获得所述第一关系网络中每个节点的向量化表示；

训练模块，被配置为通过所述训练样本数据集对分类模型进行训练，得到所述分类模型，所述分类模型使得所述样本标签与所述第一关系网络中每个节点的向量化表示相关联。

17.根据权利要求16所述的装置，其特征在于，所述第一学习模块包括：

第一生成子模块，被配置为采用随机游走算法对所述第一关系网络中每个节点进行序列采样，并生成第一节点序列；

第二学习子模块，被配置为基于预设的网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示。

18.一种数据识别装置，其特征在于，包括：

第二获取模块，被配置为获取待识别数据集，其中，所述待识别数据集包括至少两个待识别数据；

第二构建模块，被配置为构建所述至少两个待识别数据的第二关系网络，通过所述第二关系网络的节点表征所述至少两个待识别数据；

第三学习模块，被配置为基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示；

识别模块，被配置为根据预先训练的分类模型确定所述第二关系网络中每个节点的向量化表示对应的数据标签。

19.根据权利要求18所述的装置，其特征在于，所述第三学习模块包括：

第二生成子模块，被配置为采用随机游走算法对所述第二关系网络中每个节点进行序列采样，并生成第二节点序列；

第四学习子模块，被配置为根据预设的网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示。

20.一种计算设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机指令，其特征在于，所述处理器执行所述指令时实现权利要求1-7或8-15任意一项所述方法的步骤。

21.一种计算机可读存储介质，其存储有计算机指令，其特征在于，该程序被处理器执行时实现权利要求1-7或8-15任意一项所述方法的步骤。

一种分类模型生成方法及装置、一种数据识别方法及装置

技术领域

[0001] 本申请涉及计算机数据安全技术领域，特别涉及一种分类模型生成方法及装置、一种数据识别方法及装置、一种计算设备及计算机存储介质。

背景技术

[0002] 现下反洗钱行业对于犯罪等可疑交易识别的做法，主要通过人工设计业务特征，完成规则模型的构造。其中，传统的关系网络数据(例如资金网络、同设备网络等)特征，基本都是通过人工构造获得的。例如，构造网络节点或边上的统计量来刻画节点的子图模式。该类特征对于节点类别的区分判别，并没有起到很好的效果。同时，该类基于统计量的特征只能刻画节点局部(一跳)关系内子图的模式，对于多跳关系的子图，无法完整表示，有效信息会缺失。

发明内容

[0003] 有鉴于此，本申请实施例提供了一种分类模型生成方法及装置、一种数据识别方法及装置、一种计算设备及计算机存储介质，以解决现有技术中存在的技术缺陷。

[0004] 本申请实施例公开了一种分类模型生成方法，包括：

[0005] 获取训练样本数据集，其中，所述训练样本数据集中包括至少两个样本数据以及每个所述样本数据对应的样本标签；

[0006] 构建所述至少两个样本数据的第一关系网络，通过所述第一关系网络的节点表征所述至少两个样本数据；

[0007] 基于预设的网络嵌入学习模型获得所述第一关系网络中每个节点的向量化表示；

[0008] 通过所述训练样本数据集对分类模型进行训练，得到所述分类模型，所述分类模型使得所述样本标签与所述第一关系网络中每个节点的向量化表示相关联。

[0009] 另一方面，本申请实施例还提供了一种数据识别方法，包括：

[0010] 获取待识别数据集，其中，所述待识别数据集包括至少两个待识别数据；

[0011] 构建所述至少两个待识别数据的第二关系网络，通过所述第二关系网络的节点表征所述至少两个待识别数据；

[0012] 基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示；

[0013] 根据预先训练的分类模型确定所述第二关系网络中每个节点的向量化表示对应的数据标签。

[0014] 另一方面，本申请实施例还提供了一种分类模型生成装置，包括：

[0015] 第一获取模块，被配置为获取训练样本数据集，其中，所述训练样本数据集中包括至少两个样本数据以及每个所述样本数据对应的样本标签；

[0016] 第一构建模块，被配置为构建所述至少两个样本数据的第一关系网络，通过所述第一关系网络的节点表征所述至少两个样本数据；

[0017] 第一学习模块，被配置为基于预设的网络嵌入学习模型获得所述第一关系网络中

每个节点的向量化表示；

[0018] 训练模块，被配置为通过所述训练样本数据集对分类模型进行训练，得到所述分类模型，所述分类模型使得所述样本标签与所述第一关系网络中每个节点的向量化表示相关联。

[0019] 另一方面，本申请实施例还提供了一种数据识别装置，包括：

[0020] 第二获取模块，被配置为获取待识别数据集，其中，所述待识别数据集包括至少两个待识别数据；

[0021] 第二构建模块，被配置为构建所述至少两个待识别数据的第二关系网络，通过所述第二关系网络的节点表征所述至少两个待识别数据；

[0022] 第三学习模块，被配置为基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示；

[0023] 识别模块，被配置为根据预先训练的分类模型确定所述第二关系网络中每个节点的向量化表示对应的数据标签。

[0024] 另一方面，本申请还提供了一种计算设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机指令，所述处理器执行时实现所述分类模型生成方法或所述数据识别方法的步骤。

[0025] 另一方面，本申请还提供了一种计算机可读存储介质，其存储有计算机指令，该程序被处理器执行时实现所述分类模型生成方法或所述数据识别方法的步骤。

[0026] 本申请提供的一种分类模型生成方法及装置、一种数据识别方法及装置，其中，所述数据识别方法包括获取待识别数据集；构建所述至少两个待识别数据的第二关系网络，通过所述第二关系网络的节点表征所述至少两个待识别数据；基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示；根据预先训练的分类模型确定所述第二关系网络中每个节点的向量化表示对应的数据标签。

附图说明

[0027] 图1为本说明书一实施例提供的一种计算设备的结构示意图；

[0028] 图2为本说明书一实施例提供的一种分类模型生成方法的流程图；

[0029] 图3为本说明书一实施例提供的一种分类模型生成方法的流程图；

[0030] 图4为本说明书一实施例提供的一种分类模型生成方法的流程图；

[0031] 图5为本说明书一实施例提供的关系网络结构图以及关系网络结构图中每个节点的向量化表示示意图；

[0032] 图6为本说明书一实施例提供的一种数据识别方法的流程图；

[0033] 图7为本说明书一实施例提供的一种数据识别方法的流程图；

[0034] 图8为本说明书一实施例提供的一种分类模型生成装置的结构示意图；

[0035] 图9为本说明书一实施例提供的一种数据识别装置的结构示意图。

具体实施方式

[0036] 在下面的描述中阐述了很多具体细节以便于充分理解本申请。但是本申请能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本申请内涵的情况

下做类似推广，因此本申请不受下面公开的具体实施的限制。

[0037] 在本说明书一个或多个实施例中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，本说明书一个或多个实施例中使用的术语“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。

[0038] 应当理解，尽管在本说明书一个或多个实施例中可能采用术语第一、第二等来描述各种信息，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，第一也可以被称为第二，类似地，第二也可以被称为第一。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”。

[0039] 首先，对本发明一个或多个实施例涉及的名词术语进行解释。

[0040] 反洗钱：指金融机构通过流程、规则或模型等方式控制系统内的洗钱风险。

[0041] Node2vec：一种关系网络节点向量化的方法，基于Word2vec模型。

[0042] 在本说明书一个或多个实施例中，提供了一种分类模型生成方法及装置、一种数据识别方法及装置、一种计算设备及计算机存储介质，在下面的实施例中逐一进行详细说明。

[0043] 参见图1，本说明书一个实施例提供了一种计算设备100的结构框图。该计算设备100的部件包括但不限于存储器110、处理器120和通信接口140。所述处理器120与所述存储器110通过总线130相连接，数据库150用于保存训练样本数据集或待识别数据集，网络160用于生成分类模型，并与所述计算设备100通过通信接口140通信连接。

[0044] 所述通信接口140使得计算设备100能够经由一个或多个网络通信。这些网络的示例包括局域网(LAN)、广域网(WAN)、个域网(PAN)或诸如因特网的通信网络的组合。网络接口可以包括有线或无线的任何类型的网络接口(例如，网络接口卡(NIC))中的一个或多个，诸如IEEE802.11无线局域网(WLAN)无线接口、全球微波互联接入(Wi-MAX)接口、以太网接口、通用串行总线(USB)接口、蜂窝网络接口、蓝牙接口、近场通信(NFC)接口，等等。

[0045] 所述存储器110，被配置为存储通信接口140通过总线130发送的训练样本数据集或待识别数据集以及存储在存储器110上并可在处理器120上运行的计算机指令。

[0046] 所述处理器120，被配置为获取存储在存储器110的训练样本数据集或待识别数据集后，执行存储在存储器110上的计算机指令，实现对所述分类模型的生成以及数据的识别。

[0047] 所述计算设备100可以是任何类型的静止或移动计算设备，包括移动计算机或移动计算设备(例如，平板计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等)、移动电话(例如，智能手机)、可佩戴的计算设备(例如，智能手表、智能眼镜等)或其他类型的移动设备，或者诸如台式计算机或PC的静止计算设备。

[0048] 其中，处理器120可以执行图2所示方法中的步骤。图2是示出了说明书一个实施例提供的分类模型生成方法的示意性流程图，包括步骤202至步骤208。

[0049] 步骤202：获取训练样本数据集，其中，所述训练样本数据集中包括至少两个样本数据以及每个所述样本数据对应的样本标签。

[0050] 本说明书一个或多个实施例中，所述样本数据包括但不限于白样本数据和黑样本数据；所述样本标签包括但不限于白样本标签和黑样本标签。

[0051] 实际应用中，所述白样本数据以及对应的白样本标签和所述黑样本数据以及对应的黑样本标签结合在一起就构成了训练样本数据集。

[0052] 将该分类模型生成方法应用在反洗钱犯罪识别领域，所述黑样本数据可以为存在洗钱行为的非法账户，所述白样本数据可以为不存在洗钱行为的合法账户；所述黑样本数据可以通过外部舆情获取或者是通过专家经验设计业务特征获取，所述白样本数据可以通过在所有的样本中排除掉已知的黑样本数据之后获取。

[0053] 实际应用中，白样本数据的数量会远远大于黑样本数据的数量，为了避免白样本数据数量过多造成训练样本数据集中的样本数据不均衡，将白样本数据和黑样本数据的比例控制在10:1～100:1之间，因此所述白样本数据可以通过下采样获取。白样本数据和黑样本数据采用上述比例关系，可以避免导致分类模型出现过拟合，降低分类模型学习能力的负面效果的情况发生。

[0054] 步骤204：构建所述至少两个样本数据的第一关系网络，通过所述第一关系网络的节点表征所述至少两个样本数据。

[0055] 本说明书一个或多个实施例中，所述第一关系网络由一系列的节点和关系构成，节点之间会存在彼此交互等，所以节点之间出现了关系，并由此衍生出关系构建。

[0056] 本说明书一个或多个实施例中，所述至少两个样本数据代表着所述第一关系网络的节点，所述至少两个样本数据之间的关系代表着节点之间的交互。

[0057] 以所述第一关系网络为静态资金关系网络为例，对构建所述至少两个样本数据的第一关系网络进行详细说明。

[0058] 例如所述至少两个样本为获取的90天的进行过资金交易的账户，然后对所有的账户之间的资金交易情况进行汇总，若所述静态资金关系网络为有向图或有权图，则最终的静态资金关系网络呈现出三元组的形式，即：U、V、W，分别表示U到V存在权重为W的有向边，在所述静态资金关系网络对应业务场景中表示为：账户U支付给账户V金额W元。相应的，若所述第一关系网络为无向图或无权图的同设备关系网络，在所述同设备关系网络对应的业务场景中表示：账户U和账户V均采用同一设备进行资金交易，因此无需添加V到U的边，且W均设置为1即可。

[0059] 步骤206：基于预设的网络嵌入学习模型获得所述第一关系网络中每个节点的向量化表示。

[0060] 参见图3，本说明书一个或多个实施例中，基于预设的网络嵌入学习模型获得所述第一关系网络中每个节点的向量化表示包括步骤302至步骤304。

[0061] 步骤302：采用随机游走算法对所述第一关系网络中每个节点进行序列采样，并生成第一节点序列。

[0062] 步骤304：基于预设的网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示。

[0063] 本说明书一个或多个实施例中，基于预设的网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示包括：

[0064] 基于Node2vec网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化

表示;或者

[0065] 基于DeepWalk网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示。

[0066] 其中,基于Node2vec网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示包括:

[0067] 所述Node2vec网络嵌入学习模型基于Word2vec的SkipGram框架将所述节点序列中的每个节点进行向量化表示。

[0068] 本说明书一个或多个实施例中,基于Node2vec网络嵌入学习模型,采取random walk随机游走算法将构建的第一关系网络中的每个节点转化为采样后的节点序列,再进一步的基于Word2vec模型中的SkipGram框架,对采样的节点序列进行概率学习和推断,最终获得第一关系网络中每个节点向量化表示。通过网络嵌入学习获得的节点向量化表示,可以丰富节点之间的关系,提高分类模型的处理速率和效果。

[0069] 参见图4,左边为由13个节点以及各节点之间的关系构成的边构建的关系网络结构图,将所述关系网络结构图基于网络嵌入学习模型进行计算后获得右边的所述关系网络结构图中13个节点中每个节点的向量化表示,即将所述关系网络结构通过一系列过程变成一个多维向量,通过这样一层转化,能够将复杂的关系网络信息变成结构化的多维特征,从而可以利用机器学习方法实现更方便的算法应用。

[0070] 步骤208:通过所述训练样本数据集对分类模型进行训练,得到所述分类模型,所述分类模型使得所述样本标签与所述第一关系网络中每个节点的向量化表示相关联。

[0071] 本说明书一个或多个实施例中,所述分类模型包括XGBoost模型、随机森林模型、支持向量机模型SVM(Support Vector Machine)或逻辑回归模型(Logistic Regression,LR)。

[0072] 本说明书一个或多个实施例中,还可以按照预设时间间隔获取训练样本数据集,通过这种定期收集训练样本数据集、训练分类模型的方式,可以自适应的发现新特征,持续保证分类模型的有效性。

[0073] 所述预设时间间隔可以根据实际需求进行设定,比如设置为每隔一周、一个月或者三个月获取一次均可,本申请对此不作任何限定。

[0074] 本说明书一个或多个实施例中,所述分类模型生成方法只要按照预设的时间间隔定期收集数据样本,分类模型就可以自适应的学习第一关系网络特征,通过网络嵌入学习模型获取所述第一关系网络中每个节点的向量化表示,达到训练分类模型的目的,这样既能提高工作效率,又能根据所述第一关系网络中每个节点的向量化表示完整描述每个节点在第一关系网络中网络特征模式。同时通过定期收集数据样本,还可以避免训练样本数据集失效的问题。

[0075] 参见图5,本说明书一实施例提供了一种分类模型生成方法的示意性流程图,包括步骤502至步骤514。

[0076] 步骤502:获取黑样本数据以及对应的黑样本标签。

[0077] 步骤504:获取白样本数据以及对应的白样本标签。

[0078] 步骤506:将所述黑样本数据以及对应的黑样本标签和所述白样本数据以及对应的白样本标签结合形成训练样本数据集。

[0079] 步骤508:构建所述黑样本数据和所述白样本数据的关系网络,通过所述关系网络的节点表征所述黑样本数据和所述白样本数据。

[0080] 步骤510:基于Node2vec网络嵌入学习模型获得所述关系网络中每个节点的向量化表示。

[0081] 步骤512:基于所述训练样本数据集对分类模型进行训练;

[0082] 步骤514:得到所述分类模型。

[0083] 本说明书一个或多个实施例中,所述分类模型生成方法只要收集黑白数据样本,然后通过黑白样本数据构建关系网络特征,通过网络嵌入学习模型获取所述系网络中每个节点的向量化表示,达到训练分类模型的目的,这样既能提高工作效率,又能根据所述关系网络中每个节点的向量化表示完整描述每个节点在关系网络中网络特征模式。

[0084] 参见图6,本说明书一实施例提供了一种数据识别方法的示意性流程图,包括步骤602至步骤608。

[0085] 步骤602:获取待识别数据集,其中,所述待识别数据集包括至少两个待识别数据。

[0086] 本说明书一个或多个实施例中,所述待识别数据集包括通过上述分类模型判断是否存在洗钱行为的待识别账户的集合。

[0087] 步骤604:构建所述至少两个待识别数据的第二关系网络,通过所述第二关系网络的节点表征所述至少两个待识别数据。

[0088] 本说明书一个或多个实施例中,步骤604与上述实施例中步骤204的操作方式相同,在此不在赘述。

[0089] 步骤606:基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示。

[0090] 本说明书一个或多个实施例中,基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示包括:

[0091] 采用随机游走算法对所述第二关系网络中每个节点进行序列采样,并生成第二节点序列;

[0092] 根据预设的网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示。

[0093] 本说明书一个或多个实施例中,根据预设的网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示包括:

[0094] 基于Node2vec网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示;或者

[0095] 基于DeepWalk网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示。

[0096] 其中,基于Node2vec网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示包括:

[0097] 所述Node2vec网络嵌入学习模型基于Word2vec的SkipGram框架将所述第二节点序列中的每个节点进行向量化表示。

[0098] 本说明书一个或多个实施例中,将第二关系网络作为输入,通过预设的网络嵌入学习模型进行学习,得到所述第二关系网络中的每个节点的向量化表示。

[0099] 以所述预设的网络嵌入学习模型包括Node2vec网络嵌入学习模型为例，对通过网络嵌入学习模型得到所述第二关系网络中的每个节点的向量化表示进行详细的说明。

[0100] 第一步：基于第二关系网络结构，计算第二关系网络中每条边的转移概率，获得第二关系网络的转移概率图。

[0101] 例如，第二关系网络有节点U、节点V和节点K，其中出边(有向图)权重之和为Z，每条出边的转移概率为：P(V)＝W(U,V)/Z，其中W(U,V)表示节点U到节点V的边权重。通过上述方式即可获得第二关系网络的转移概率图。

[0102] 第二步：基于第二关系网络的转移概率图随机游走生成第二关系网络中的每个节点的节点序列。

[0103] 本说明书一个或多个实施例中，随机游走构造出第二关系网络中的每个节点的节点序列应当满足如下约束条件：假定当前节点V，上一次随机游走节点为U，V的邻居节点K，如果K和U的最短路径距离为0，则转移概率为W(U,V)/Z/p，其中p为模型参数；如果K和U的最短路径距离为1，则转移概率为W(U,V)/Z；如果K和U的最短路径距离为2，则转移概率为W(U,V)/Z/q，其中q为模型参数。重复以上随机转移过程并满足以上约束条件，直到序列长度达到指定参数MAX_LEN停止，其中MAX_LEN为模型参数。

[0104] 第三步：所述Node2vec网络嵌入学习模型基于Word2vec的SkipGram框架将所述节点序列中的每个节点进行向量化表示。

[0105] 本说明书一个或多个实施例中，使用Word2vec的SkipGram框架采用随机梯度下降法进行模型的优化学习，最终获得第二关系网络中每个节点的向量化表示。

[0106] 步骤608：根据预先训练的分类模型确定所述第二关系网络中每个节点的向量化表示对应的数据标签。

[0107] 本说明书一个或多个实施例中，所述数据标签包括白数据标签和黑数据标签。

[0108] 实际应用中，所述白数据标签对应的为不存在洗钱行为的合法账户，所述黑样本数据标签对应的为存在洗钱行为的非法账户。若所述第二关系网络中每个节点的向量化表示对应的数据标签为白样本数据标签，则该节点的向量化表示对应的待识别数据为合法账户；若所述第二关系网络中每个节点的向量化表示对应的数据标签为黑样本数据标签，则该节点的向量化表示对应的待识别数据为非法账户。

[0109] 本说明书一个或多个实施例中，若所述数据标签为黑数据标签，则对所述数据标签对应的所述第二关系网络中节点的向量化表示进行后续处理。

[0110] 所述后续处理包括但不限于进行账户资金流转追溯或者是账户对应的真实用户的详细身份查询以及登陆信息查询，本申请对此不作任何限定。

[0111] 本说明书一个或多个实施例中，所述数据识别方法通过根据待识别数据集构建第二关系网络，可以较为完整的描述该第二关系网络的局部子图模式，然后根据预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示，通过预先训练的分类模型快速的确定所述第二关系网络中每个节点的向量化表示对应的数据标签，实现对待识别数据的快速识别。

[0112] 参见图7，本说明书一实施例提供了一种数据识别方法的示意性流程图，包括步骤702至步骤710。

[0113] 步骤702：获取至少两个待识别账户。

[0114] 步骤704:将所述两个待识别账户形成待识别账户集。

[0115] 步骤706:构建所述待识别账户集的关系网络,通过所述关系网络的节点表征所述待识别账户集。

[0116] 步骤708:基于Node2vec网络嵌入学习模型获得所述关系网络中每个节点的向量化表示。

[0117] 步骤710:根据预先训练的分类模型确定所述关系网络中每个节点的向量化表示对应的账户标签。

[0118] 本说明书一个或多个实施例中,将该方法应用在反洗钱领域,使用关系网络这类原始信息作为输入,通过网络嵌入学习模型及预先训练的分类模型,实现对犯罪行为账户的快速识别。

[0119] 参见图8,本说明书一实施例提供了一种分类模型生成装置,包括:

[0120] 第一获取模块802,被配置为获取训练样本数据集,其中,所述训练样本数据集中包括至少两个样本数据以及每个所述样本数据对应的样本标签;

[0121] 第一构建模块804,被配置为构建所述至少两个样本数据的第一关系网络,通过所述第一关系网络的节点表征所述至少两个样本数据;

[0122] 第一学习模块806,被配置为基于预设的网络嵌入学习模型获得所述第一关系网络中每个节点的向量化表示;

[0123] 训练模块808,被配置为通过所述训练样本数据集对分类模型进行训练,得到所述分类模型,所述分类模型使得所述样本标签与所述第一关系网络中每个节点的向量化表示相关联。

[0124] 可选地,所述第一学习模块806包括:

[0125] 第一生成子模块,被配置为采用随机游走算法对所述第一关系网络中每个节点进行序列采样,并生成第一节点序列;

[0126] 第二学习子模块,被配置为基于预设的网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示。

[0127] 本说明书一个或多个实施例中,所述分类模型生成装置只要按照预设的时间间隔定期收集数据样本,分类模型就可以自适应的学习第一关系网络特征,通过网络嵌入学习模型获取所述第一关系网络中每个节点的向量化表示,达到训练分类模型的目的,这样既能提高工作效率,又能根据所述第一关系网络中每个节点的向量化表示完整描述每个节点在第一关系网络中网络特征模式。同时通过定期收集数据样本,还可以避免训练样本数据集失效的问题。

[0128] 参见图9,本说明书一实施例提供了一种数据识别装置,包括:

[0129] 第二获取模块902,被配置为获取待识别数据集,其中,所述待识别数据集包括至少两个待识别数据;

[0130] 第二构建模块904,被配置为构建所述至少两个待识别数据的第二关系网络,通过所述第二关系网络的节点表征所述至少两个待识别数据;

[0131] 第三学习模块906,被配置为基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示;

[0132] 识别模块908,被配置为根据预先训练的分类模型确定所述第二关系网络中每个

节点的向量化表示对应的数据标签。

[0133] 可选地，所述第三学习模块906包括：

[0134] 第二生成子模块，被配置为采用随机游走算法对所述第二关系网络中每个节点进行序列采样，并生成第二节点序列；

[0135] 第四学习子模块，被配置为根据预设的网络嵌入学习模型将所述第二节点序列中的每个节点进行向量化表示。

[0136] 本说明书一个或多个实施例中，所述数据识别装置通过根据待识别数据集构建第二关系网络，可以较为完整的描述该第二关系网络的局部子图模式，然后根据预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示，通过预先训练的分类模型快速的确定所述第二关系网络中每个节点的向量化表示对应的数据标签，实现对待识别数据的快速识别。

[0137] 本说明书一个或多个实施例中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于装置实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0138] 本申请一实施例还提供一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时实现所述分类模型生成方法的步骤。

[0139] 上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是，该存储介质的技术方案与上述的分类模型生成方法的技术方案属于同一构思，存储介质的技术方案未详细描述的细节内容，均可以参见上述分类模型生成方法的技术方案的描述。

[0140] 本申请一实施例还提供一种计算机可读存储介质，其存储有计算机指令，该指令被处理器执行时实现所述数据识别方法的步骤。

[0141] 上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是，该存储介质的技术方案与上述的数据识别方法的技术方案属于同一构思，存储介质的技术方案未详细描述的细节内容，均可以参见上述数据识别方法的技术方案的描述。

[0142] 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。

[0143] 本申请实施例中所述支付涉及的技术载体，例如可以包括近场通信(Near Field Communication，NFC)、WIFI、3G/4G/5G、POS机刷卡技术、二维码扫码技术、条形码扫码技术、蓝牙、红外、短消息(Short Message Service，SMS)、多媒体消息(Multimedia Message Service，MMS)等。

[0144] 所述计算机指令包括计算机指令代码，所述计算机指令代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括：能够携带所述计算机指令代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是，所述计算机可读介质包含的内

容可以根据司法管辖区内立法和专利实践的要求进行适当的增减，例如在某些司法管辖区，根据立法和专利实践，计算机可读介质不包括电载波信号和电信信号。

[0145] 需要说明的是，对于前述的各方法实施例，为了简便描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定都是本申请所必须的。

[0146] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其它实施例的相关描述。

[0147] 以上公开的本申请优选实施例只是用于帮助阐述本申请。可选实施例并没有详尽叙述所有的细节，也不限制该发明仅为所述的具体实施方式。显然，根据本说明书的内容，可作很多的修改和变化。本说明书选取并具体描述这些实施例，是为了更好地解释本申请的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本申请。本申请仅受权利要求书及其全部范围和等效物的限制。

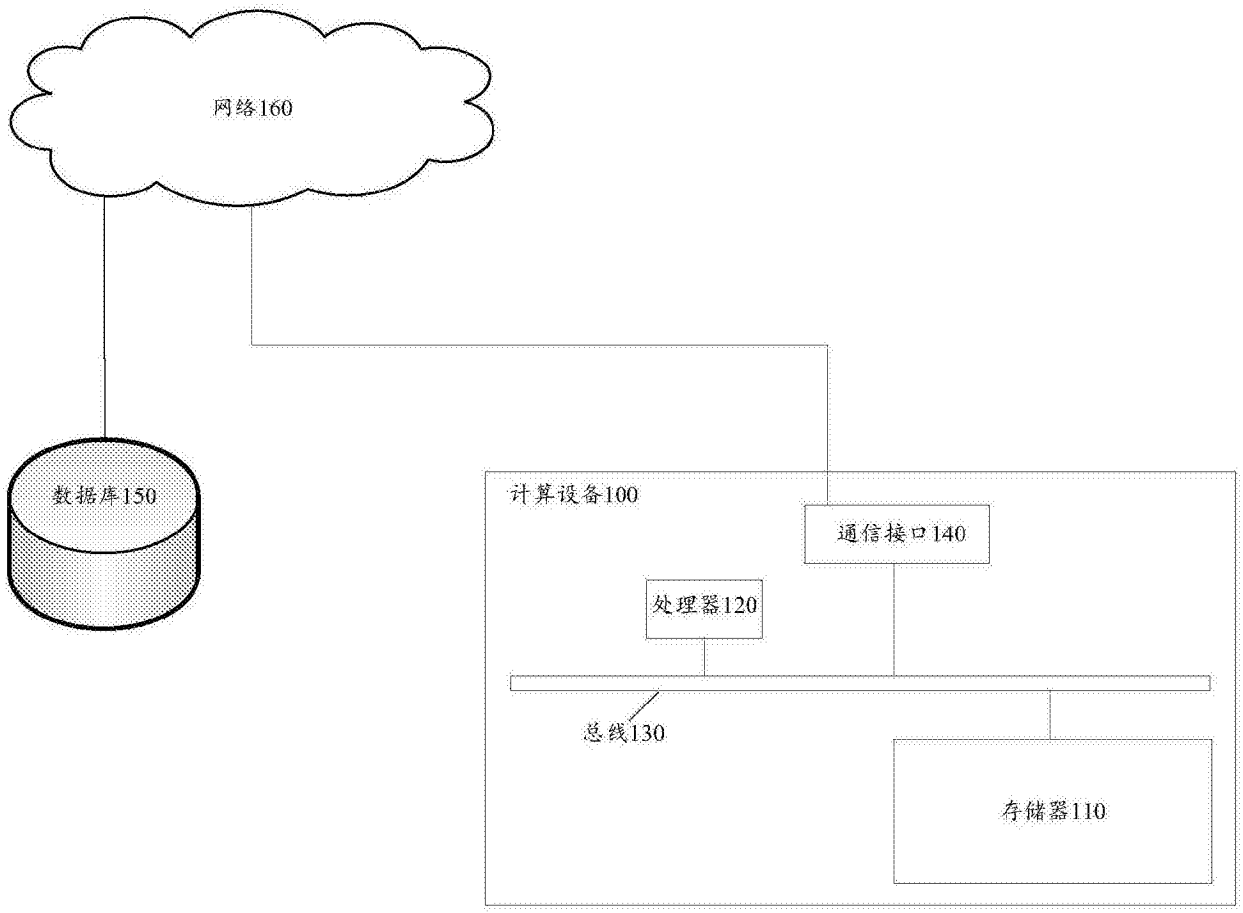


图1

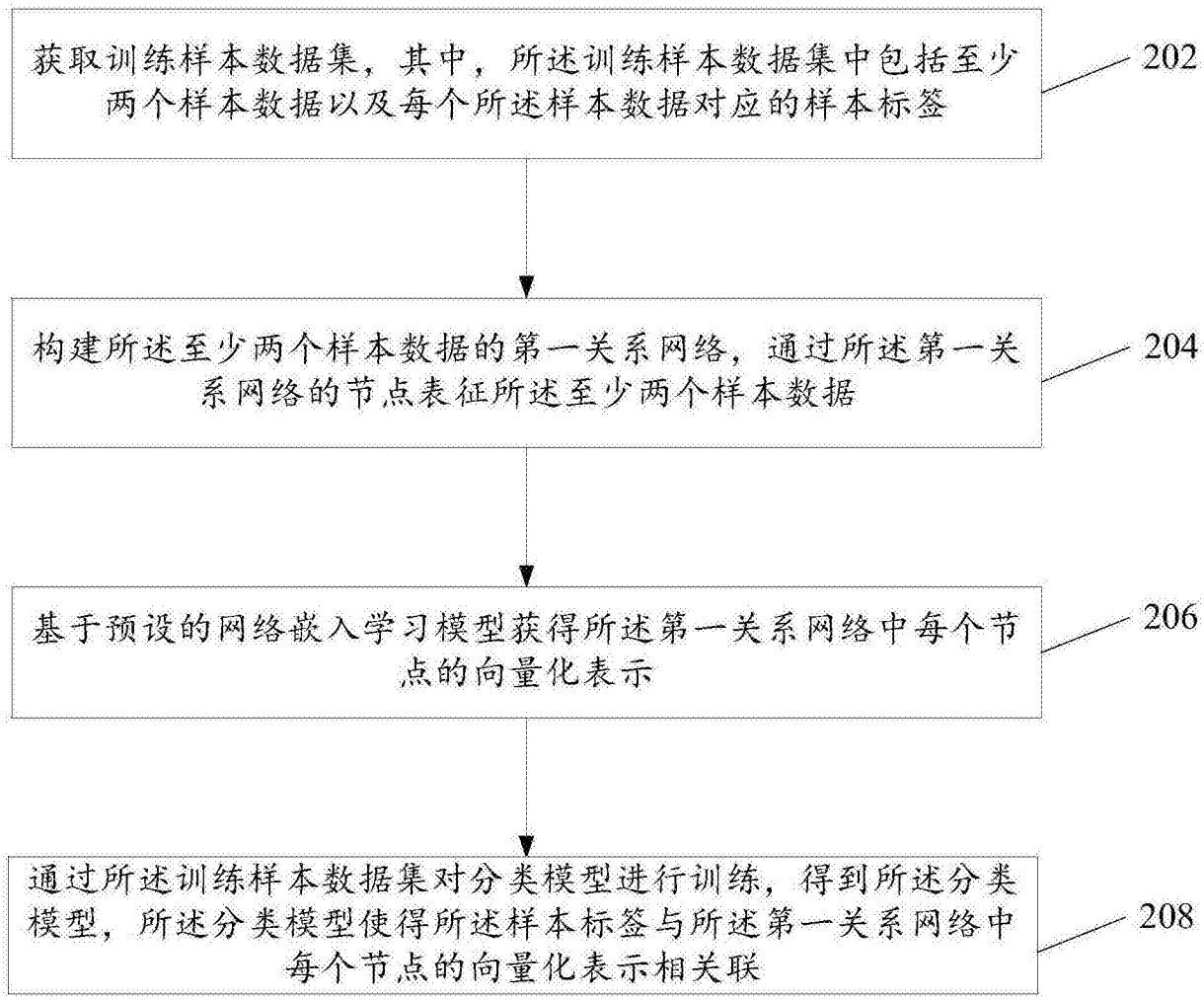

获取训练样本数据集，其中，所述训练样本数据集中包括至少两个样本数据以及每个所述样本数据对应的样本标签
202
构建所述至少两个样本数据的第一关系网络，通过所述第一关系网络的节点表征所述至少两个样本数据
204
基于预设的网络嵌入学习模型获得所述第一关系网络中每个节点的向量化表示
206
通过所述训练样本数据集对分类模型进行训练，得到所述分类模型，所述分类模型使得所述样本标签与所述第一关系网络中每个节点的向量化表示相关联
208

图2

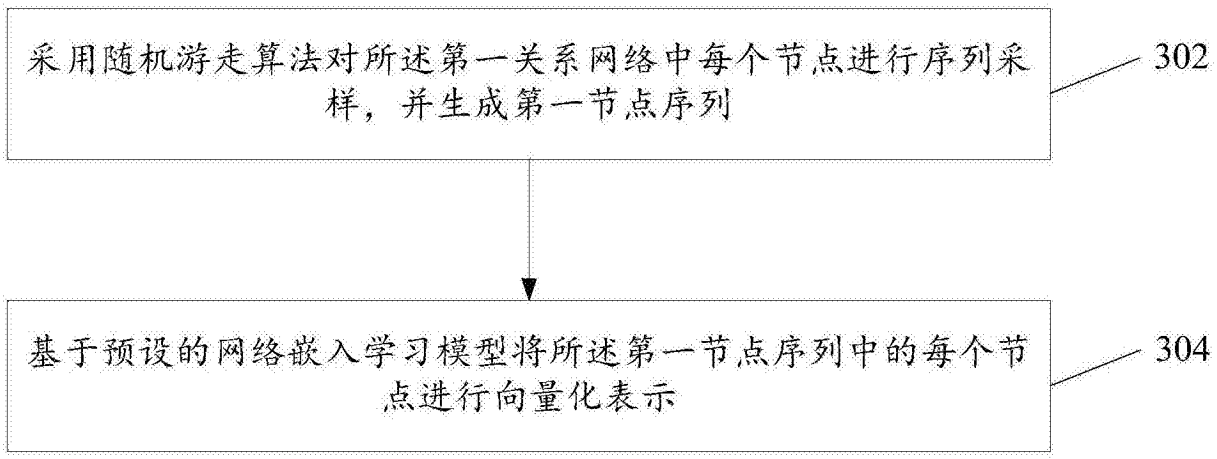

采用随机游走算法对所述第一关系网络中每个节点进行序列采样，并生成第一节点序列
302
基于预设的网络嵌入学习模型将所述第一节点序列中的每个节点进行向量化表示
304

图3

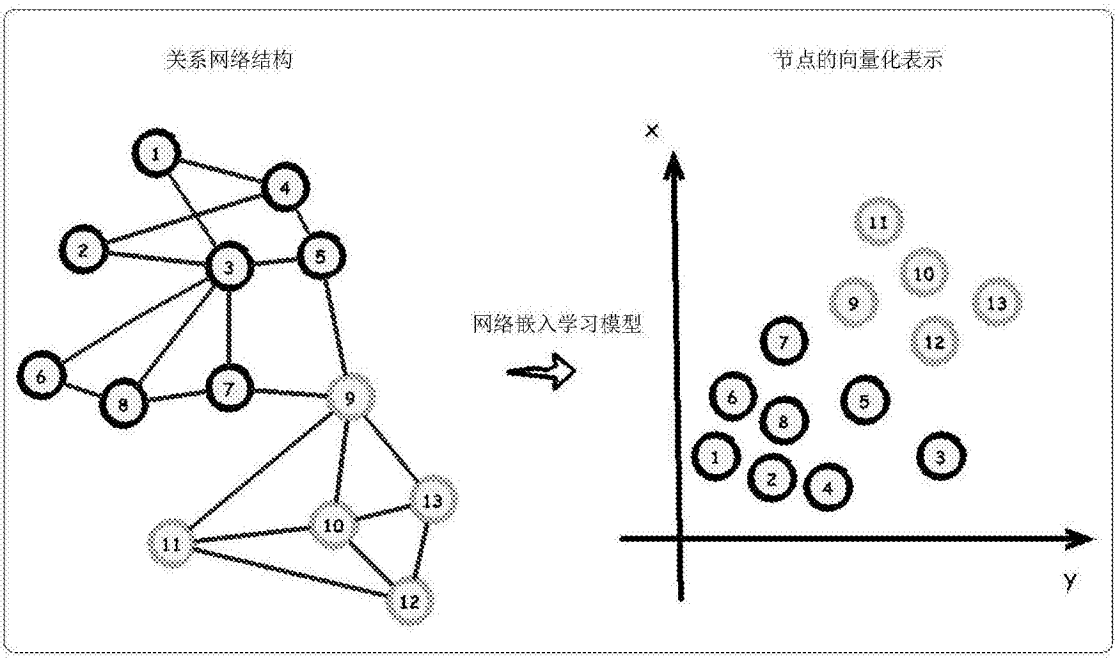


图4

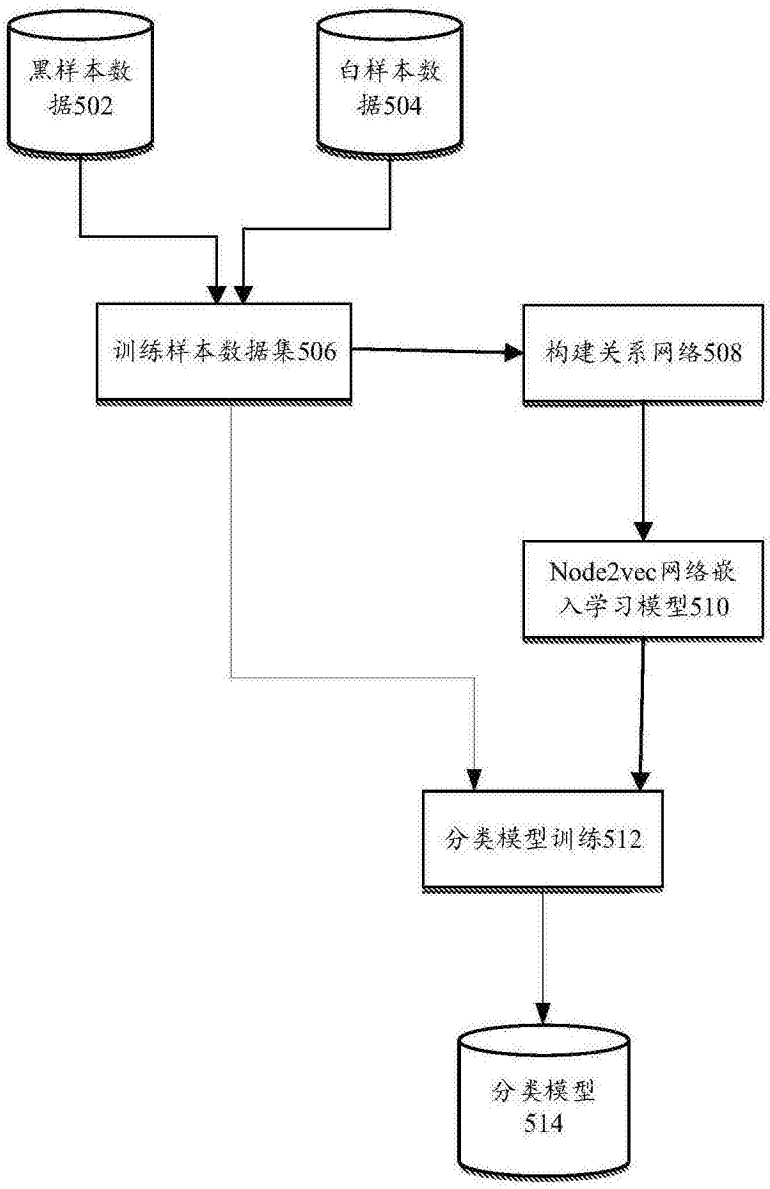

黑样本数据502
白样本数据504
训练样本数据集506
构建关系网络508
Node2vec网络嵌入学习模型510
分类模型训练512
分类模型514

图5

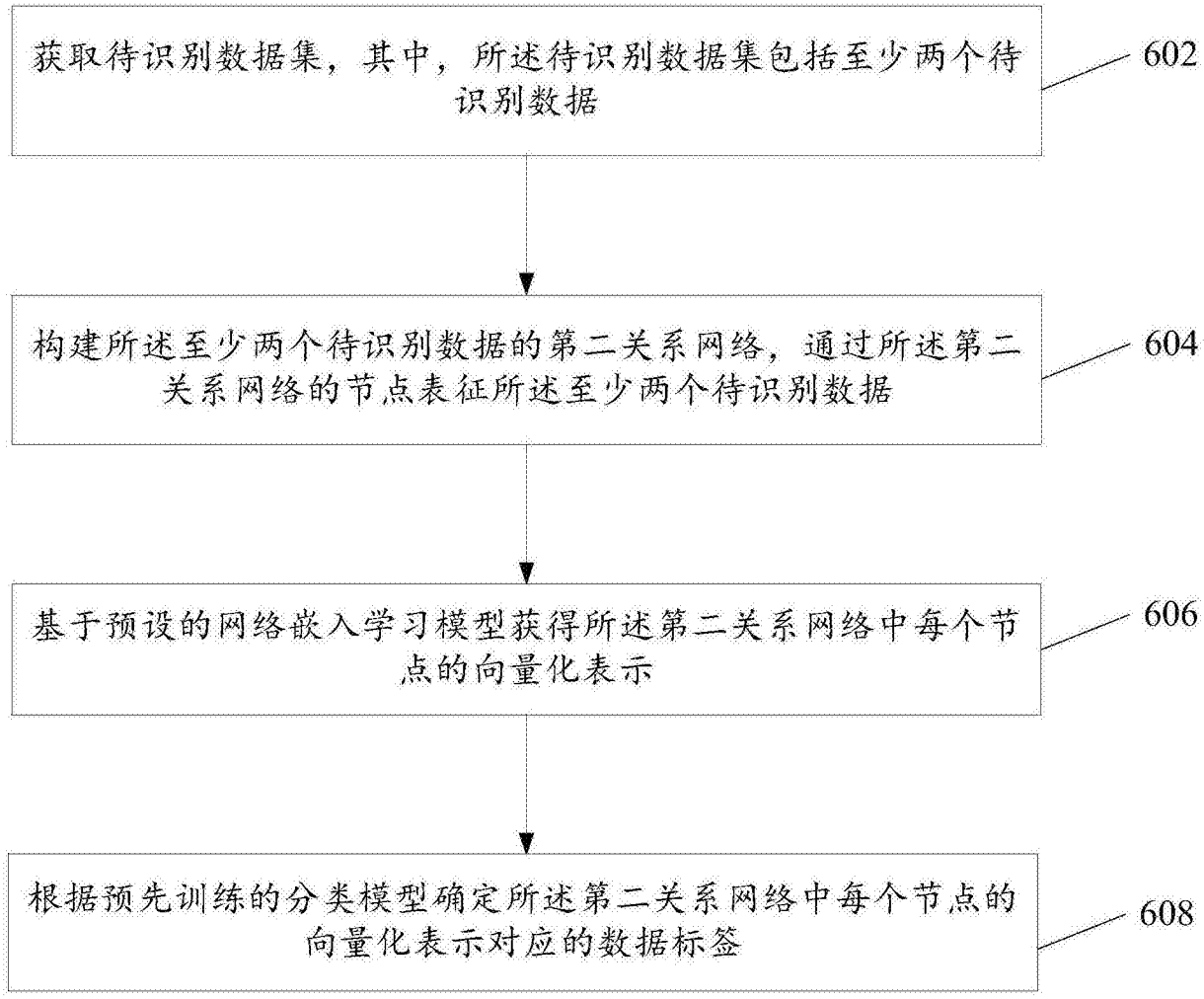

获取待识别数据集，其中，所述待识别数据集包括至少两个待识别数据
602
构建所述至少两个待识别数据的第二关系网络，通过所述第二关系网络的节点表征所述至少两个待识别数据
604
基于预设的网络嵌入学习模型获得所述第二关系网络中每个节点的向量化表示
606
根据预先训练的分类模型确定所述第二关系网络中每个节点的向量化表示对应的数据标签
608

图6

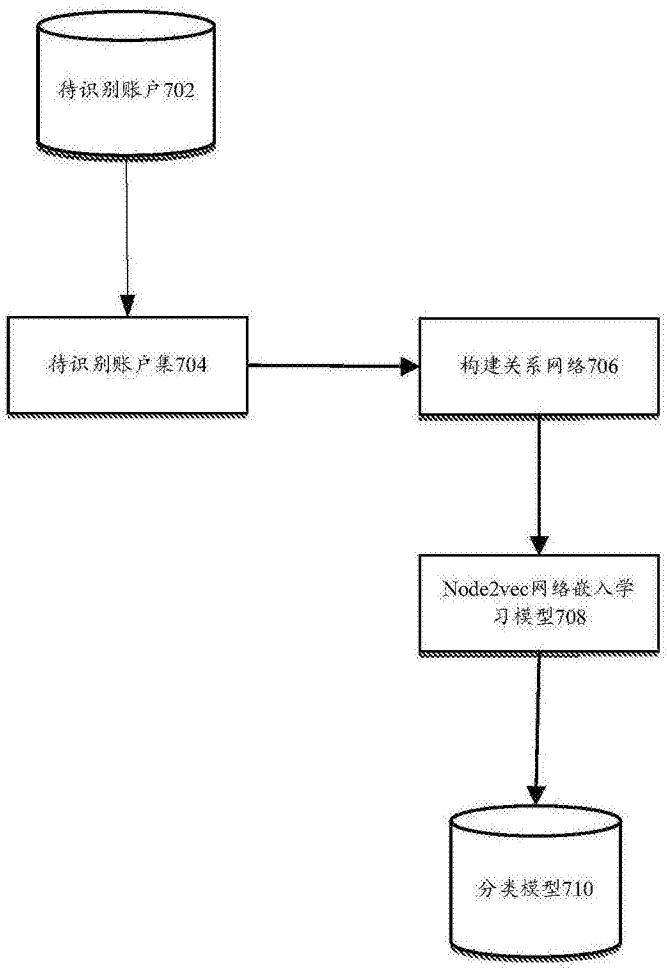

待识别账户702
待识别账户集704
构建关系网络706
Node2vec网络嵌入学习模型708
分类模型710

图7

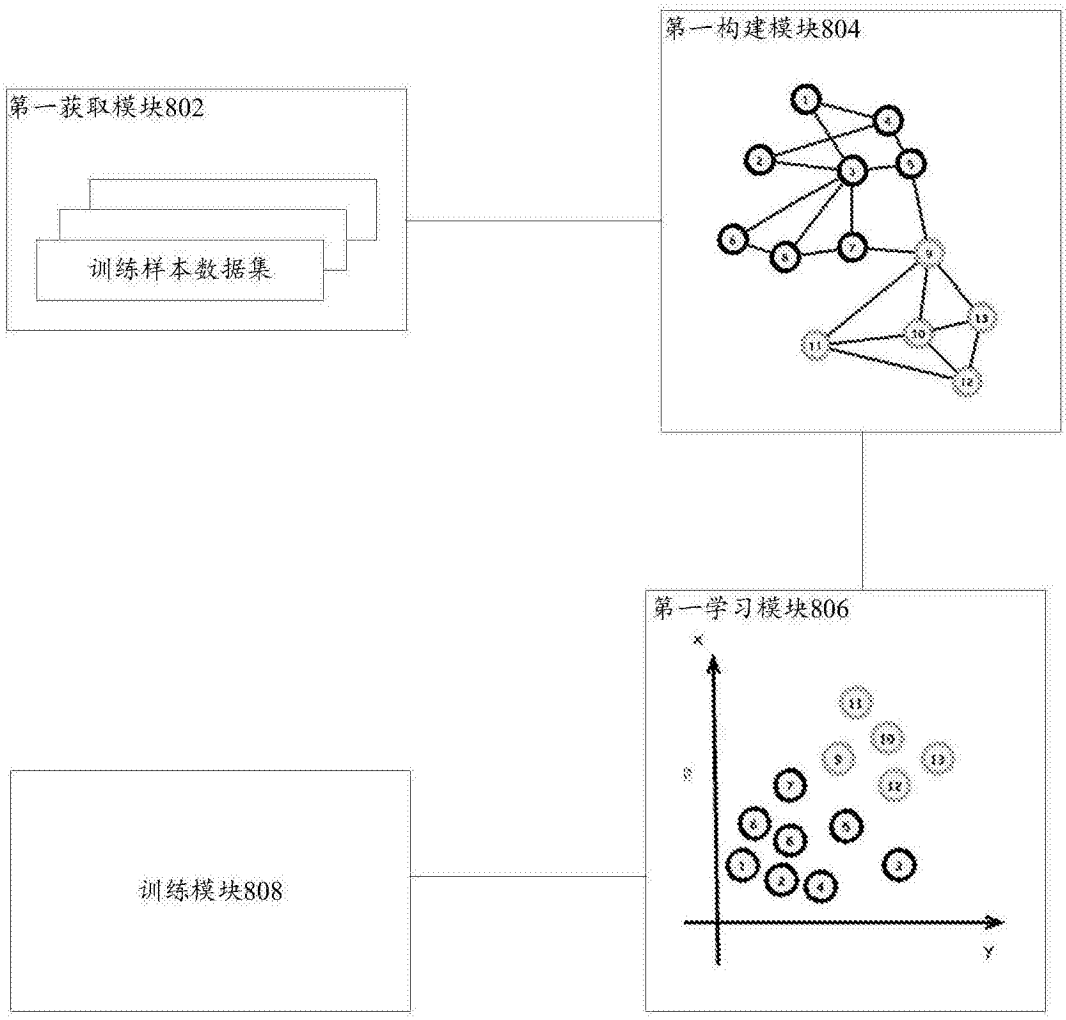


图8

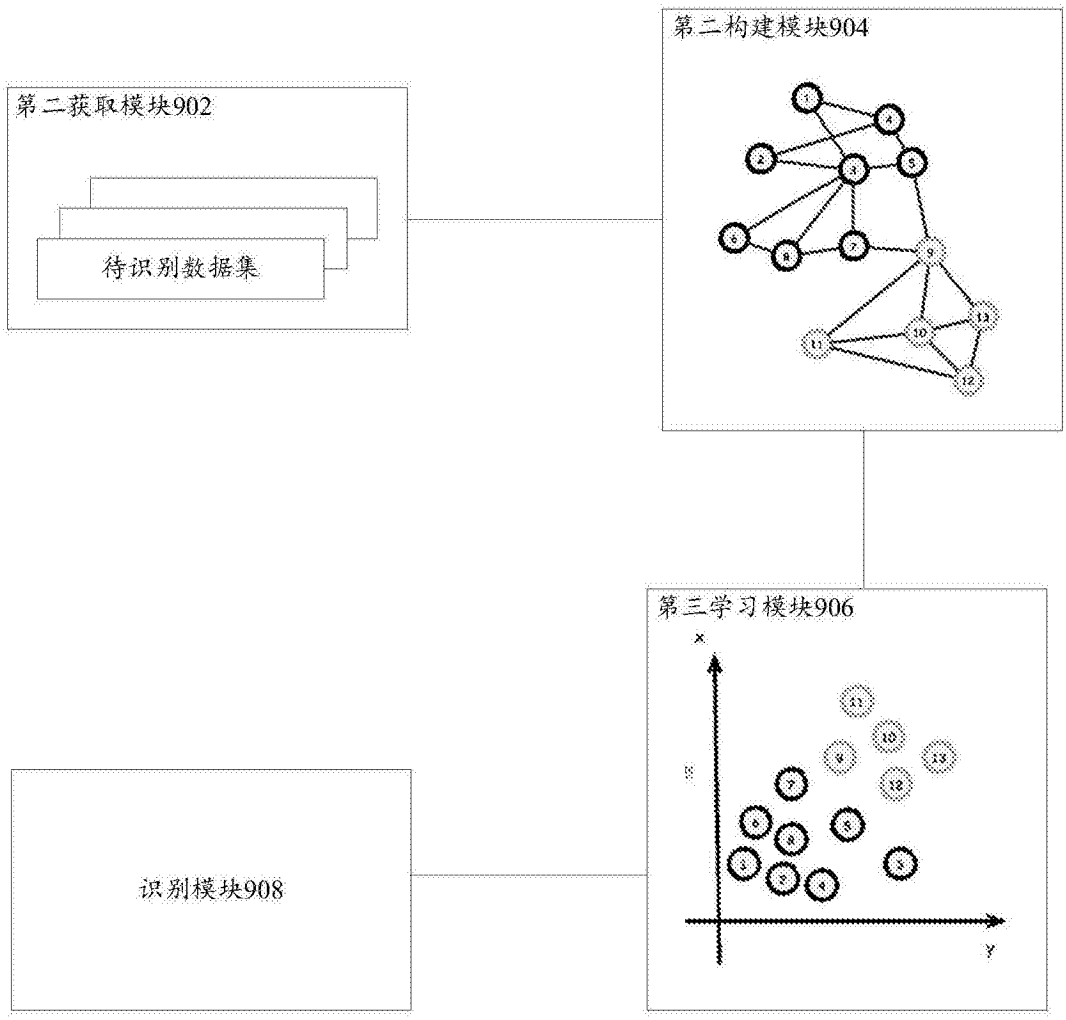


图9