

序列挖掘模型的训练方法、序列数据的处理方法及设备

本申请要求于 2020 年 02 月 18 日提交的申请号为 202010099547.7、发明名称为“序列挖掘模型的训练方法、序列数据的处理方法及设备”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及人工智能和机器学习技术领域，特别涉及一种序列挖掘模型的训练方法、序列数据的处理方法及设备。

背景技术

诸如信用卡反欺诈、反小号作弊、反薅羊毛党等风控业务场景中，通常会面临海量的序列数据，对海量的序列数据进行处理，以挖掘和区分“好”与“坏”的长期行为模式，在风控业务场景中尤为重要。

相关技术针对风控业务场景中序列数据的挖掘，提供了一种有监督学习方案。首先，利用 DTW (Dynamic Time Warping, 动态时间归整) 算法对所有序列样本进行规整对齐，再计算两两序列样本间的相似度，根据序列间相似度，利用 KNN (K-Nearest Neighbor, K 最邻近) 模型对所有的序列样本进行分类，从而可以将“好”的序列样本与“坏”的序列样本进行区分。

然而，由于不同的风控业务场景中，序列样本并非都是具备标签的，相关技术提供的技术方案仅适用于某一能够提供具备标签的序列样本的风控业务场景，不具备通用性。

发明内容

本申请实施例提供了一种序列挖掘模型的训练方法、序列数据的处理方法及设备，可以充分利用业务场景中的历史序列数据，提升资源利用率，并且提升通过序列挖掘模型确定序列数据的标签信息的准确性。所述技术方案如下：

一方面，本申请实施例提供了一种序列挖掘模型的训练方法，应用于计算机设备中，所述方法包括：

获取目标业务场景下的第一序列样本，所述第一序列样本包括所述目标业务场景下的历史序列数据；

确定所述第一序列样本的标签状态，所述第一序列样本的标签状态用于指示所述第一序列样本具备标签信息的情况；

根据所述标签状态，从序列挖掘框架中选取子模型构建序列挖掘模型，所述序列挖掘模型用于确定所述目标业务场景下序列数据的标签信息；

采用所述第一序列样本对所述序列挖掘模型进行训练；

其中，所述序列挖掘框架包括第一子模型、第二子模型和第三子模型；所述第一子模型用于获取所述序列数据的隐向量表示；所述第二子模型用于在所述标签状态满足第一条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息；所述第三子模型用于在所述标签状态满

足第二条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

另一方面，本申请实施例提供了一种序列数据的处理方法，应用于计算机设备中，所述方法包括：

获取目标业务场景的序列数据；

调用所述目标业务场景对应的序列挖掘模型，所述序列挖掘模型是根据所述目标业务场景下的历史序列数据的标签状态，从序列挖掘框架中选取子模型构建并采用所述历史序列数据进行训练得到的模型；

通过所述序列挖掘模型对所述序列数据进行处理，得到所述序列数据的标签信息；

其中，所述序列挖掘框架包括第一子模型、第二子模型和第三子模型；所述第一子模型用于获取所述序列数据的隐向量表示；所述第二子模型用于在所述标签状态满足第一条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息；所述第三子模型用于在所述标签状态满足第二条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

再一方面，本申请实施例提供了一种序列挖掘模型的训练装置，所述装置包括：

样本获取模块，用于获取目标业务场景下的第一序列样本，所述第一序列样本包括所述目标业务场景下的历史序列数据；

状态确定模块，用于确定所述第一序列样本的标签状态，所述第一序列样本的标签状态用于指示所述第一序列样本具备标签信息的情况；

模型构建模块，用于根据所述标签状态，从序列挖掘框架中选取子模型构建序列挖掘模型，所述序列挖掘模型用于确定所述目标业务场景下序列数据的标签信息；

模型训练模块，用于采用所述第一序列样本对所述序列挖掘模型进行训练；

其中，所述序列挖掘框架包括第一子模型、第二子模型和第三子模型；所述第一子模型用于获取所述序列数据的隐向量表示；所述第二子模型用于在所述标签状态满足第一条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息；所述第三子模型用于在所述标签状态满足第二条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

又一方面，本申请实施例提供了一种序列数据的处理装置，所述装置包括：

数据获取模块，用于获取目标业务场景的序列数据；

模型调用模块，用于调用所述目标业务场景对应的序列挖掘模型，所述序列挖掘模型是根据所述目标业务场景下的历史序列数据的标签状态，从序列挖掘框架中选取子模型构建并采用所述历史序列数据进行训练得到的模型；

标签确定模块，用于通过所述序列挖掘模型对所述序列数据进行处理，得到所述序列数据的标签信息；

其中，所述序列挖掘框架包括第一子模型、第二子模型和第三子模型；所述第一子模型用于获取所述序列数据的隐向量表示；所述第二子模型用于在所述标签状态满足第一条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息；所述第三子模型用于在所述标签状态满

足第二条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

还一方面，本申请实施例提供了一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现上述序列挖掘模型的训练方法，或者实现上述序列数据的处理方法。

还一方面，本申请实施例提供了一种计算机可读存储介质，所述存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现上述序列挖掘模型的训练方法，或者实现上述序列数据的处理方法。

还一方面，本申请实施例提供了一种计算机程序产品，当所述计算机程序产品在计算机设备上运行时，使得计算机设备执行上述序列挖掘模型的训练方法，或者执行上述序列数据的处理方法。

本申请实施例提供的技术方案，提出了一种序列挖掘框架，实际应用中通过获取目标业务场景下的序列样本，根据该序列样本的标签状态从序列挖掘框架中选取子模型构建序列挖掘模型，然后采用该序列样本对序列挖掘模型进行训练，从而使得训练出的序列挖掘模型可以用来判断目标业务场景下序列数据的标签信息。并且，本申请实施例提出的序列挖掘框架可以在不同的业务场景中通用，该序列挖掘框架中包括至少一个子模型，针对不同的业务场景下历史序列数据不同的标签状态，可以从序列挖掘框架中选取对应的子模型构建序列挖掘模型。另外，相比于相关技术中只能采用结构固定的单一模型进行训练，可能造成某一业务场景下部分历史序列数据无法使用，如相关技术中只能采用无监督聚类模型进行训练，可能造成某一业务场景下存在标签信息的历史序列数据无法使用，产生资源浪费的问题，本申请实施例提供的技术方案，针对历史序列数据不同的标签状态，可以采用不同的子模型进行训练，从而充分利用了业务场景中的所有历史序列数据，提升了资源利用率。此外，由于本申请实施例提供的技术方案中，可以根据目标业务场景下所有的历史序列数据充分训练序列挖掘模型，从而提升了通过该序列挖掘模型确定序列数据的标签信息的准确性。

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例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申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1是本申请一个实施例提供的实施环境的示意图；

图2是本申请一个实施例提供的序列挖掘模型的训练方法的流程图；

图3是本申请另一个实施例提供的序列挖掘模型的训练方法的流程图；

图4是本申请一个实施例提供的预训练第一子模型的方法的流程图；

图5是本申请一个实施例提供的第一子模型的框图；

图6是本申请一个实施例提供的优化向量表示的计算方法的流程图；

图7是本申请一个实施例提供的联合训练方法的流程图；

图8是本申请再一个实施例提供的序列挖掘模型的训练方法的流程图；

图 9 是本申请又一个实施例提供的序列挖掘模型的训练方法的流程图，

图 10 是本申请一个实施例提供的序列数据的处理方法的流程图；

图 11 是本申请一个实施例提供的序列挖掘模型的训练方法和序列数据的处理方法的流程图；

图 12 是本申请一个实施例提供的序列挖掘框架的示意图；

图 13 是本申请一个实施例提供的序列挖掘模型的训练装置的框图；

图 14 是本申请另一个实施例提供的序列挖掘模型的训练装置的框图；

图 15 是本申请一个实施例提供的序列数据的处理装置的框图；

图 16 是本申请一个实施例提供的计算机设备的结构框图。

具体实施方式

为使本申请的目的、技术方案和优点更加清楚，下面将结合附图对本申请实施方式作进一步地详细描述。

人工智能（Artificial Intelligence，AI）是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、推理与决策的功能。

人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

机器学习（Machine Learning，ML）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、示教学习等技术。

本申请实施例提供的技术方案涉及人工智能的机器学习/深度学习技术，具体通过如下实施例进行说明。

请参考图 1，其示出了本申请一个实施例提供的实施环境的实施例。该实施环境可以包括：第一计算机设备 10 和第二计算机设备 20。

计算机设备是指具备数据分析处理功能和存储功能的设备，如具备数据存储能力和/或数据处理能力的 PC（Personal Computer，个人计算机）和服务器等，或者是诸如手机、平板电脑、多媒体播放设备、可穿戴设备等终端，还可以是其他计算机设备，本申请实施例对此不作限定。可选地，当计算机设备为服务器时，该计算机设备可以是一台服务器，也可以是由多台服务器组成的服务器集群，或者是一个云计算服务中心。

第一计算机设备 10 是指具体业务场景中具备数据存储和传输功能的计算机设备。可选地，一种具体业务场景可以对应一台第一计算机设备 10，可以对应多台第一计算机设备 10。

本申请实施例中，第一计算机设备 10 可以获取其所应用的具体业务场景，将序列数据进行存储，或者将序列数据发送给第二计算机设备 20，由第二计算机设备 20 对序列数据进行分析处理。

第二计算机设备 20 具备数据分析处理功能，能够对序列数据进行处理。可选地，第二计算机设备 20 可以根据序列数据训练机器学习模型，和/或调用机器学习模型判断序列数据的标签信息。可选地，第二计算机设备 20 中存储有序列挖掘框架，针对某一具体业务场景，第二计算机设备 20 可以从该序列挖掘框架中选取部分或全部子模型组成对应于该具体业务场景的序列挖掘模型，并采用该具体业务场景下的历史序列数据对该序列挖掘模型进行训练，以使得训练后的序列挖掘模型可以用来判断该具体业务场景下的序列数据的标签信息。

本申请实施例中，第一计算机设备 10 和第二计算机设备 20 之间可以通过网络进行互相通信。该网络可以是有线网络，也可以是无网络。示例性地，第二计算机设备 20 可以从第一计算机设备 10 处获取目标业务场景中的历史序列数据，并根据这些历史序列数据组成序列样本，然后根据该序列样本的标签信息，从序列挖掘框架中选取对应的子模型组成序列挖掘模型，接着使用序列样本对该序列挖掘模型进行训练，以使得第二计算机设备 20 可以通过调用该序列挖掘模型，对该目标业务场景中的序列数据进行处理，得到该序列数据的标签信息。

请参考图 2，其示出了本申请一个实施例提供的序列挖掘模型的训练方法的流程图。该方法可应用于上文介绍的计算机设备中。该方法可以包括如下几个步骤（210~240）：

步骤 210，获取目标业务场景下的第一序列样本。

目标业务场景是指需要确定其业务场景中序列数据的标签信息的业务场景，本申请实施例对目标业务场景的类型不作限定，可选地，目标业务场景包括金融交易业务场景（如证券交易系统和货币支付系统）、物联网业务场景（如可穿戴设备和智能家居系统）、环境监测业务场景（如湿度监测系统和空气颗粒物监测系统）、医疗智能业务场景（如身体关键指标跟踪系统和健康状况监测系统）和事件应用程序应用场景（如用户信息管理系统和会员管理系统）等。序列数据，又称为时间序列数据，是指目标业务场景在不同时间上收集到的数据，用于描述目标业务场景中出现的现象随时间变化的情况。例如，假设目标业务场景为会员管理系统，则用户在该会员管理系统对应的平台注册会员账号后，其在该平台的每一项行为都会形成序列数据，对应于该用户的所有序列数据可以反映该用户在该平台注册会员后的行为模式，如反映该用户使用会员资格购买折扣商品的频率。

第一序列样本是指用于训练序列挖掘模型的样本，本申请实施例中，第一序列样本包括目标业务场景下的历史序列数据。可选地，第一序列样本中包含至少一个序列样本，每个序列样本既可以由一条历史序列数据组成，也可以由多条历史序列数据组成，本申请实施例对此不作限定，实际应用中可以结合计算机设备的处理开销和实际业务场景确定第一序列样本中序列样本的个数，以及序列样本的具体组成。

步骤 220，确定第一序列样本的标签状态。

第一序列样本的标签状态用于指示第一序列样本具备标签信息的情况。其中，标签信息用于反映第一序列样本的特征属性，本申请实施例中，标签信息的具体内容根据目标业务场景确定，可选地，在对序列数据的标签信息准确性要求较低的情况下，为了减少模型训练时间和降低计算机设备的处理开销，标签信息可以包括“好”和“坏”两类，其中，“好”是指与目标业务场景的主体业务相吻合，“坏”是指与目标业务场景的主体业务相背离。例如，假设目

标业务场景为信用卡管理系统，那么在该目标业务场景下，第一序列样本的标签信息可以包括“好”和“坏”，其中，“好”是指正常使用信用卡，“坏”是指异常使用信用卡，若该目标业务场景对序列数据的标签信息准确性要求较高，那么可以将标签信息进一步细分，如将“坏”细分为“洗钱”和“诈骗”等。

由于序列数据在生成时不一定具备标签信息，所以在不同的业务场景中，序列数据具备标签信息的情况可能不相同，因而不同的业务场景对应的第一序列样本的标签状态也可能是不一样的。本申请实施例对目标业务场景中第一序列样本的标签状态不作限定，可选地，在目标业务场景中，第一序列样本的标签状态可以是第一序列样本全部不存在标签信息，也可以是第一序列样本部分存在标签信息，部分不存在标签信息。

步骤 230，根据标签状态，从序列挖掘框架中选取子模型构建序列挖掘模型。

序列挖掘框架用于针对不同的业务场景构建序列挖掘模型，本申请实施例提供的序列挖掘框架中包括多个子模型，计算机设备可以根据第一序列样本的标签状态，即目标业务场景中历史序列数据具备标签信息的情况，从序列挖掘框架中选取至少一个子模型构建序列挖掘模型。本申请实施例通过提出融合了多个子模型的序列挖掘框架，使得针对不同的业务场景，均可以使用该序列挖掘框架构建与其业务场景中序列数据的标签状态相适应的序列挖掘模型，相比于相关技术中只能采用结构固定的单一模型对序列数据进行处理，可能造成资源浪费的问题，例如，只能采用无监督聚类模型对序列数据进行处理，此时，业务场景中存在标签信息的序列样本就无法使用该模型，本申请实施例提供的技术方案，可以充分利用业务场景中的所有历史序列数据，提升了资源利用率。

可选地，序列挖掘框架包括第一子模型、第二子模型和第三子模型。其中，第一子模型用于获取序列数据的隐向量表示，序列数据的隐向量表示是指提取了序列数据中的关键特征信息后构成的向量表示，可选地，第一子模型可以是深度自编码模型（Autoencoder，AE）。有关隐向量表示的具体获取过程，请参见下述可选实施例，此处不再赘述。第二子模型用于在标签状态满足第一条件的情况下，根据序列数据的隐向量表示确定序列数据的标签信息，第一条件包括标签状态为第一序列样本不存在标签信息，可选地，第二子模型可以是聚类模型，如 K-means 聚类模型。第三子模型用于在标签状态满足第二条件的情况下，根据序列数据的隐向量表示确定序列数据的标签信息，第二条件包括标签状态为第一序列样本至少部分存在标签信息，可选地，第三子模型可以是标准分类模型、PU Learning（Learning from Positive and Unlabeled Example，正例和无标注样本学习）模型等。有关第一子模型、第二子模型和第三子模型的组成结构和实现方式，请参见下述可选实施例，此处不再赘述。

步骤 240，采用第一序列样本对序列挖掘模型进行训练。

序列挖掘模型用于确定目标业务场景下序列数据的标签信息。计算机设备根据第一序列样本的标签状态，即目标业务场景中的历史序列数据的标签状态，从序列挖掘框架中选取与标签状态相适应的子模型构建序列挖掘模型后，可以采用第一序列样本对序列挖掘模型进行训练，从而使得训练之后的序列挖掘模型可以用来确定目标业务场景中序列数据的标签信息。本申请实施例提供的技术方案可以采用目标业务场景下所有的历史序列数据对序列挖掘模型进行训练，相比于相关技术采用结构固定的单一模型时，可能只能采用与该单一模型相适应的部分历史序列数据对该单一模型进行训练，本申请实施例提供的技术方案根据目标业务场景下历史序列数据充分训练了序列挖掘模型，提升了通过该序列挖掘模型确定序列数据的标签信息的准确性。

可选地，第二子模型和第三子模型也可以同时作为构建某一序列挖掘模型的子模型，此时，它们在序列挖掘模型中所起的作用可能不同。例如，假设在目标业务场景中第一序列样本部分存在标签信息，部分不存在标签信息，此时可以采用第三子模型对存在标签信息的第一序列样本进行训练，训练第三子模型的目的是使得第一子模型获取的隐向量表示更为可靠；采用第二子模型对不存在标签信息的第一序列样本进行训练，训练第二子模型的目的是将第一序列样本中不存在标签信息的序列样本按照目标业务场景下可能的标签信息种类进行分类，从而使得第二子模型可以学习不同标签信息的序列样本表现出的行为模式，进而在目标业务场景中有新的序列数据到来时，可以最终通过该第三子模型确定序列数据的标签信息。有关第二子模型和第三子模型同时作为构建某一序列挖掘模型的子模型时，第二子模型和第三子模型训练过程的具体示例，请参见下述可选实施例，此处不再赘述。

需要说明的一点是，本申请实施例仅以采用目标业务场景下所有的历史序列数据组成第一序列样本为例，对有益效果和序列挖掘模型的训练过程进行举例说明，本领域技术人员在了解了本申请实施例的技术方案后，将很容易想到其它的技术方案，如实际应用中考虑到计算机设备的处理开销和样本的代表性等因素，可以对目标业务场景下的海量历史序列数据进行筛选等处理，以选取部分历史序列数据组成第一序列样本，但仍应属于本申请实施例的保护范围。

综上所述，本申请实施例提供的技术方案，提出了一种序列挖掘框架，实际应用中通过获取目标业务场景下的序列样本，根据该序列样本的标签状态从序列挖掘框架中选取子模型构建序列挖掘模型，然后采用该序列样本对序列挖掘模型进行训练，从而使得训练出的序列挖掘模型可以用来判断目标业务场景下序列数据的标签信息。并且，本申请实施例提出的序列挖掘框架可以在不同的业务场景中通用，该序列挖掘框架中包括至少一个子模型，针对不同的业务场景下历史序列数据不同的标签状态，可以从序列挖掘框架中选取对应的子模型构建序列挖掘模型。另外，相比于相关技术中只能采用结构固定的单一模型进行训练，可能造成某一业务场景下部分历史序列数据无法使用，如相关技术中只能采用无监督聚类模型进行训练，可能造成某一业务场景下存在标签信息的历史序列数据无法使用，产生资源浪费的问题，本申请实施例提供的技术方案，针对历史序列数据不同的标签状态，可以采用不同的子模型进行训练，从而充分利用了业务场景中的所有历史序列数据，提升了资源利用率。此外，由于本申请实施例提供的技术方案中，可以根据目标业务场景下所有的历史序列数据充分训练序列挖掘模型，从而提升了通过该序列挖掘模型确定序列数据的标签信息的准确性。

在一种可能的实施方式中，上述标签状态为第一序列样本不存在标签信息，上述序列挖掘模型包括第一子模型和第二子模型；如图 3 所示，上述步骤 240 包括如下几个子步骤（242~246）：

步骤 242，采用第一序列样本对第一子模型进行预训练，得到预训练的第一子模型。

第一子模型是用于获取序列数据的隐向量表示的子模型，可选地，第一子模型可以是深度自编码模型，其可以通过将第一序列样本作为学习目标，对第一序列样本进行表征学习（representation learning），以得到第一序列样本的隐向量表示（Latent Representation）。本申请实施例中，计算机设备采用第一序列样本对第一子模型进行预训练，以使得通过预训练的第一子模型可以得到第一序列样本较为可靠的隐向量表示。并且，由于第一序列样本的隐向量表示相比于第一序列样本，保留了第一序列样本中关键特征信息的同时，降低了第一序列

样本的维度和序列长度，从而可以在采用第一序列样本的隐向量表示训练、推理、测试时，降低计算机设备的处理开销。

在一个示例中，上述第一子模型包括隐向量提取单元和重构单元；如图 4 所示，上述步骤 242 包括如下几个子步骤（2422~2428）：

步骤 2422，通过所述隐向量提取单元对所述第一序列样本进行处理，得到所述第一序列样本的隐向量表示。

如图 5 所示，其示出了一种可能的第一子模型的框图，该第一子模型包括隐向量提取单元 410 和重构单元 420，在第一子模型实现为深度自编码模型时，隐向量提取单元 410 可以实现为深度自编码模型中的编码器，重构单元 420 可以实现为深度自编码模型中的解码器。其中，计算机设备可以通过隐向量提取单元 410 对第一序列样本进行处理，以得到第一序列样本的隐向量表示。

可选地，如图 5 所示，上述隐向量提取单元 410 包括映射子单元 412、降维子单元 414 和递归子单元 416；上述步骤 2422 包括如下几个子步骤：

（1）通过所述映射子单元获取所述第一序列样本的特征信息，得到所述第一序列样本的特征信息表示。

可选地，本申请实施例中，第一子模型输入的第一序列样本的形状可以表示为：`[batch_size,time_steps,feature_nums]`，其中，“`batch_size`”用于表示第一序列样本中的序列样本数量，“`time_steps`”用于表示第一序列样本的长度，“`feature_nums`”用于表示第一序列样本的特征个数。

计算机设备可以通过映射子单元对第一序列样本的特征信息进行提取，同时保留第一序列样本的序列结构。在第一子单元实现为深度自编码模型时，映射子单元可以实现为 1DCNN（One Dimensions Convolutional Neural Network，一维深度卷积神经网络），计算机设备通过 1DCNN 对第一序列样本进行从上向下的卷积核映射，获取第一序列样本的特征信息。

（2）通过降维子单元对第一序列样本的特征信息表示进行降维处理，得到第一序列样本的降维特征信息。

在获取到第一序列样本的特征信息后，可以进一步对第一序列样本进行压缩，以避免处理较长的序列数据导致计算机设备处理开销较大。第一子模型中的降维子单元可以对第一序列样本的特征信息表示的维度进行降低，得到第一序列样本的降维特征信息。在第一子模型实现为深度自编码模型时，降维子单元可以实现为 max pooling（最大池化）操作，在 max pooling 操作期间可以使用 Leaky rectifying 激活函数，计算机设备通过 max pooling 操作将第一序列样本压缩为紧凑的向量表示，即第一序列样本的降维特征信息。

（3）通过递归子单元学习第一序列样本的降维特征信息的序列间跨步长的变化规律，得到第一序列样本的隐向量表示。

为了进一步压缩第一序列样本，计算机设备可以通过递归子单元学习第一序列样本的降维特征信息之间的序列间跨步长变化规律，从而得到第一序列样本的隐向量表示，完成对第一序列样本的编码步骤。在第一子模型实现为深度自编码模型时，递归子单元可以实现为双向 LSTM（Long Short-Term Memory，长短期记忆）网络，该双向 LSTM 网络可以从两个方向（正向和反向）学习序列间跨步长的变化规律，从而对第一序列样本进行进一步压缩，得到第一序列样本的隐向量表示。

步骤 2424，通过重构单元对第一序列样本的隐向量表示进行处理，得到重构后的第一序

列样本。

本申请实施例中，为了使得通过第一子模型得到的隐向量表示较为可靠，计算机设备在通过隐向量提取单元获取到第一序列样本的隐向量表示后，可以通过重构单元对该第一序列样本的隐向量表示进行处理，以将该隐向量表示还原为序列数据，得到重构后的第一序列样本。在第一子模型实现为深度自编码模型时，重构单元可以实现为 UpSample（升采样函数）和 Deconvolution（逆卷积）操作，计算机设备将第一序列样本的隐向量表示通过 UpSample 后，再进行 Deconvolution 操作，可以得到一个输出序列，从而完成对隐向量表示的解码过程，其中，该输出序列即为重构后的第一序列样本。

步骤 2426，根据第一序列样本和重构后的第一序列样本，计算第一损失函数的值。

第一损失函数是指第一子模型对应的损失函数。根据第一序列样本和重构后的第一序列样本，计算机设备可以确定第一损失函数，从而可以计算第一损失函数的值，可选地，第一损失函数的值可以是重构后的第一序列样本的平均绝对误差（Mean Absolute Error, MAE）。即第一损失函数的值可以通过如下表达式计算得到：

$$ae_loss = \frac{1}{n} \sum_{i=1}^n |h(x_i) - x_i|$$

其中， ae_loss 表示第一损失函数的值， x_i 表示第一序列样本中的第 i 个输入的序列样本， $h(x_i)$ 表示经过第一子模型的隐向量提取单元和重构单元后与 x_i 对应的重构后的序列样本， i 为正整数。

步骤 2428，根据第一损失函数的值，调整第一子模型的参数，得到预训练的第一子模型。

在模型的训练过程中，可以通过调整该模型的参数使得该模型的损失函数收敛，以完成对该模型的训练。本申请实施例中，计算机设备根据第一损失函数的值，调整第一子模型的参数，使得第一损失函数收敛，从而得到预训练的第一子模型。

步骤 244，通过预训练的第一子模型对第一序列样本进行处理，得到第一序列样本的隐向量表示。

预训练的第一子模型由于损失函数收敛，重构后的第一序列样本与原始输入的第一序列样本之间的误差进入容忍度范围，因而通过预训练的第一子模型获得的第一序列样本的隐向量表示较为可靠。计算机设备在得到预训练的第一子模型后，通过该预训练的第一子模型再次对第一序列样本进行处理，得到第一序列样本的隐向量表示。

可选地，本申请实施例中，第一序列样本的隐向量表示的形状可以表示为：[batch_size, time_step2, 1]，其中，“time_step2”表示第一序列样本的隐向量表示的序列长度，“1”表示第一序列样本的隐向量表示的特征个数为 1。本申请实施例中，“time_step2”满足“time_step2 < time_step”，“time_step”表示第一序列样本的序列长度。

在一个示例中，如图 6 所示，上述步骤 244 之后还可以包括如下几个步骤（2442~2444）：

步骤 2442，获取第一序列样本的静态特征。

在实际应用中，目标业务场景中还包括一些静态数据，该静态数据反映序列数据稳定的静态特征。例如，假设目标业务场景为信用卡管理系统，则序列数据可以包括信用卡用户历史的交易金额序列和历史交易时间间隔序列等，静态数据可以包括信用卡用户注册电话号码数量、所在国家、所属国籍等，相关技术中仅考虑第一序列样本的序列数据，而忽略了第一序列样本的静态特征，从而使得这些静态数据在模型训练过程中没有产生作用，训练出的模型不够准确。本申请实施例提供的技术方案中，计算机设备可以通过获取第一序列样本的静

态特征,使得这些静态数据可以应用于模型的训练过程中,提升了序列挖掘模型的训练数据资源的利用率。

步骤 2444,根据第一序列样本的静态特征和第一序列样本的隐向量表示,得到第一序列样本的优化向量表示。

计算机设备可以将第一序列样本的静态特征结合到第一序列样本的隐向量表示中,以得到第一序列样本的优化向量表示,该第一序列样本的优化向量表示用于后续对预训练的第一子模型和第二子模型进行联合训练。可选地,对于第一序列样本中的第 i 个序列样本,其隐向量表示记为 z_i , z_i 的形状可以表示为: $[\text{time_step2}, 1]$ 。假设第 i 个序列样本具有 n 个静态特征, n 为正整数,则第 i 个序列样本的静态特征的形状可以表示为: $[n, 1]$ 。计算机设备根据第 i 个序列样本的隐向量表示和静态特征,可以得到第 i 个序列样本的优化向量表示,可选地,第 i 个序列样本的优化向量表示的形状可以表示为: $[n+\text{time_step2}, 1]$ 。

可选地,上述步骤 2444 包括如下几个子步骤:

(1) 对所述第一序列样本的静态特征进行正则化处理,得到正则化处理后的静态特征。

正则化处理是指保留数据所有的特征变量的同时,减小特征变量的数量级,以避免过拟合问题的出现,即由于特征变量的数量过多,造成模型过于复杂的问题。本申请实施例中,由于第一序列样本包含至少一个静态特征,因此可以预先对第一序列样本的静态特征进行正则化处理,以减小第一序列样本的静态特征的数量级,得到正则化处理后的静态特征。

(2) 对所述第一序列样本的隐向量表示进行泛化处理,得到泛化处理后的隐向量表示。

本申请实施例中,计算机设备可以对第一序列样本的隐向量表示进行泛化处理,以调整第一序列样本的隐向量表示的数量级。可选地,计算机设备进行泛化处理可以通过在第一子模型的递归单元中加入泛化器来实现,例如,在第一子模型实现为深度自编码模型时,递归单元可以是双向 LSTM 网络,此时,可以在双向 LSTM 网络中加入 Layer Normalization (层泛化),以调整隐向量表示的数量级。

(3) 结合正则化处理后的静态特征和泛化处理后的隐向量表示,得到第一序列样本的优化向量表示。

通过对第一序列样本的静态特征进行正则化处理,以及对第一序列样本的隐向量表示进行泛化处理,可以使得正则化处理后的静态特征的数值大小和泛化处理后的隐向量表示的数值大小位于相同的数量级上。然后结合该正则化处理后的静态特征和泛化处理后的隐向量表示,得到第一序列样本的优化向量表示,该进行了数量级调整的优化向量表示,相比于直接将静态特征和隐向量表示相结合的优化向量表示,可以降低计算机设备进行后续联合训练过程的处理开销和提升计算精度。

步骤 246,采用第一序列样本和第一序列样本的隐向量表示,对预训练的第一子模型和第二子模型进行联合训练,得到序列挖掘模型。

计算机设备可以采用第一序列样本和第一序列样本的隐向量表示,对预训练的第一子模型和第二子模型进行联合训练,即采用第一序列样本对预训练的第一子模型进行训练,同时采用第一序列样本的隐向量表示对第二子模型进行训练,从而得到序列挖掘模型,该序列挖掘模型包括联合训练的第一子模型和联合训练的第二子模型。

在一个示例中,如图 7 所示,上述步骤 246 包括如下几个子步骤 (2462~2464):

步骤 2462,计算第一损失函数的值和第二损失函数的值。

其中,第一损失函数是指第一子模型的损失函数,第二损失函数是指第二子模型的损失

函数。有关第一损失函数的值的具体计算过程，请参见上述图 5 所示可运行代码，此处不再赘述。可选地，上述计算第二损失函数的值，包括如下几个步骤：

(1) 确定第一序列样本的 k 个类别。

对于目标业务场景而言，序列数据具备的标签信息可以划分为多个类别，根据标签信息的类别，可以确定第一序列样本的类别数量 k，k 为正整数。可选地，第一序列样本的类别数量，可以等于标签信息的类别数量，例如，在目标业务场景中，序列数据的标签信息具有两个类别：“好”和“坏”，那么计算机设备根据标签信息的类别，可以确定第一序列样本也可以划分为两个类别。

(2) 根据 k 个类别，通过第二子模型初始化 k 个聚类中心。

计算机设备根据 k 个类别，可以通过第二子模型在第一序列样本中随机初始化 k 个聚类中心，然后通过 k-means (k 均值) 聚类算法可以更新聚类中心。本申请实施例中，第二子模型的参数即为聚类中心，其形状可以表示为参数矩阵：[k, time_step2, 1]。

(3) 计算第一序列样本至 k 个聚类中心的距离，得到样本概率分布。

样本概率分布用于指示第一序列样本分别属于 k 个类别的概率。计算机设备初始化 k 个聚类中心后，可以计算第一序列样本中每一个序列样本分别属于这 k 个聚类中心的概率，即计算第一序列样本中每一个序列样本距离各个聚类中心的距离。根据计算出的第一序列样本至 k 个聚类中心的距离，可以看出一个概率分布，本申请实施例中，将该概率分布记为样本概率分布。

(4) 根据样本概率分布计算仿真样本概率分布。

仿真样本概率分布用于指示在第一序列样本存在标签信息的情况下，第一序列样本分别属于 k 个类别的概率。由于第一序列样本不存在标签信息，因此计算机设备无法得知真实的样本概率分布，此时，可以由样本概率分布得到一个仿真样本概率分布，以该仿真样本概率分布无限接近真实的样本概率分布。可选地，仿真样本概率分布的表达式可以如下所示：

$$p_{i,j} = \frac{q_{i,j}^2/f_i}{\sum_{j=1}^k q_{i,j}^2/f_i}, \quad f_i = \sum_{i=1}^n q_i$$

其中， $p_{i,j}$ 表示第一序列样本中第 i 个序列样本属于第 j 个聚类中心的仿真概率，所有的 $p_{i,j}$ 组合在一起即形成仿真样本概率分布； $q_{i,j}$ 表示第一序列样本中第 i 个序列样本属于第 j 个聚类中心的概率，所有的 $q_{i,j}$ 组合在一起即形成样本概率分布； f_i 表示样本概率分布。

(5) 将样本概率分布和仿真样本概率分布的相对熵确定为第二损失函数的值。

相对熵又称为 KL 散度 (Kullback-Leibler divergence, KLD)、信息散度和信息增益，是两个概率分布差别的非对称的度量。本申请实施例中，样本概率分布和仿真样本概率分布的相对熵，是用来度量使用基于仿真样本概率分布的编码来编码基于样本概率分布的样本平均所需的额外的比特个数，计算机设备以样本概率分布和仿真样本概率分布的相对熵，作为第二损失函数的值。可选地，第二损失函数的值的表达式如下所示：

$$cluster_loss = \sum_{i=1}^n \sum_{j=1}^k p_{i,j} \log \frac{p_{i,j}}{q_{i,j}}$$

其中， $cluster_loss$ 表示第二损失函数的值。

步骤 2464，根据第一损失函数的值、第二损失函数的值和权重系数设置，计算序列挖掘模型的损失函数的值。

权重系数设置是指对第一损失函数的值的权重和第二损失函数的值的权重的设置，可选

的, 权重系数设置可以由计算机设备预先设定, 也可以在序列挖掘模型的训练过程中, 根据目标应用场景和实际需求具体确定, 本申请实施例对权重系数设置的确定时机不作限定。计算机设备根据第一损失函数的值、第二损失函数的值和权重系数设置, 可以计算序列挖掘模型的损失函数的值, 可选地, 序列挖掘模型的损失函数的值可以表示如下:

$$loss_{\text{联合训练}} = r_1 * ae_loss + r_2 * cluster_loss$$

其中, $loss_{\text{联合训练}}$ 表示序列挖掘模型的损失函数的值, 即预训练的第一子模型和第二子模型进行联合训练时损失函数的值; r_1 和 r_2 表示权重系数设置。

步骤 2466, 根据序列挖掘模型的损失函数的值, 调整序列挖掘模型的参数, 得到序列挖掘模型。

本申请实施例中, 计算机设备根据序列挖掘模型的损失函数的值, 调整序列挖掘模型的参数, 以使得序列挖掘模型的损失函数收敛, 从而得到完成联合训练的序列挖掘模型。

综上所述, 本申请实施例提供的技术方案, 在标签状态为序列样本不存在标签信息时, 确定序列挖掘模型包括第一子模型和第二子模型。并且, 本申请实施例在对序列挖掘模型进行训练时, 首先对第一子模型进行预训练, 得到预训练的第一子模型, 由于预训练的第一子模型的损失函数收敛, 从而序列样本经过预训练的第一子模型得到的隐向量表示较为可靠; 接着, 采用序列样本及其隐向量表示对预训练的第一子模型和第二子模型进行联合训练, 得到序列挖掘模型, 从而提供了一种在序列样本不存在标签信息时, 序列挖掘模型的具体训练方法。

另外, 本申请实施例提供的技术方案, 可以获取序列样本的静态特征, 并将序列样本的静态特征与序列样本的隐向量表示结合, 从而得到序列样本的优化向量表示, 然后使用该序列样本及其优化向量表示, 对预训练的第一子模型和第二子模型进行联合训练, 得到序列挖掘模型, 相比于相关技术中仅考虑序列样本的序列数据, 忽略序列样本的静态特征, 使得静态数据在模型训练过程中没有产生作用, 训练出的模型不够准确, 本申请实施例提供的技术方案充分考虑了序列样本的特征数据, 使得静态特征也可以应用于模型的训练过程中, 提升了序列挖掘模型的准确性和数据资源的利用率。

此外, 本申请实施例提供的技术方案在结合序列样本的静态特征与隐向量表示时, 对静态特征进行正则化处理、对隐向量进行泛化处理, 从而使得静态特征与隐向量表示位于相同的数量级, 使得降低计算机设备采用优化后的向量表示训练序列挖掘模型时的处理开销, 并且提升计算精度。

在另一种可能的实施方式中, 上述标签状态为第一序列样本至少部分存在标签信息, 且第二序列样本的数量大于第三序列样本的数量; 其中, 第二序列样本是指第一序列样本中不存在标签信息的序列样本; 第三序列样本是指第一序列样本中不存在标签信息的序列样本; 上述序列挖掘模型包括第一子模型和第三子模型; 如图 8 所示, 上述步骤 240 包括如下几个子步骤 (24A~24E):

步骤 24A, 采用第一序列样本对第一子模型进行预训练, 得到预训练的第一子模型。

有关对第一子模型进行预训练过程的详细介绍说明, 请参见上述可选实施例, 此处不再赘述。需要说明的一点是, 本申请实施例中也可以采用第二序列样本或者第三序列样本对第

一子模型进行预训练，得到预训练的第一子模型，从而可以提升计算机设备的预训练速度，降低计算机设备的处理开销。

步骤 24C，通过预训练的第一子模型对第二序列样本进行处理，得到第二序列样本的隐向量表示。

由于第一序列样本中存在标签信息的第二序列样本的数量较多，因而考虑到序列挖掘模型的训练速度和计算机设备的处理开销，可以通过训练第三子模型，以实现最终可以通过第三子模型确定序列数据的标签信息。本申请实施例中，第三子模型是在序列样本中至少部分存在标签信息的情况下，可以通过序列数据的隐向量表示确定序列数据的标签信息的模型，且第三子模型的训练是采用第二序列样本，即第一序列样本中存在标签信息的序列样本，进行训练的，实际应用中，第三子模型可以实现为有监督学习模型，如分类模型。

由于目前已有的可以实现为第三子模型的分类模型多用于处理二维的序列数据，因此，计算机设备在得到预训练的第一子模型后，可以通过该预训练的第一子模型对第二序列样本进行处理，以得到第二序列样本的隐向量表示，从而在第三子模型的训练过程中，采用第二序列样本中的关键特征信息对第三子模型进行训练，避免采用三维的第二序列样本对第三子模型进行训练，一方面可以利用目前已有的分类器作为第三子模型，另一方面可以降低计算机设备训练第三子模型时的处理开销。

可选地，为了提升序列挖掘模型的准确性和数据资源的利用率，计算机设备在得到第二序列样本的隐向量表示之后，还包括：获取第二序列样本的静态特征；根据第二序列样本的静态特征和第二序列样本的隐向量表示，得到第二序列样本的优化向量表示；其中，第二序列样本的优化向量表示用于对预训练的第一子模型和第三子模型进行联合训练。有关第二序列样本的静态特征及其获取过程、第二序列样本的优化向量表示的计算过程的详细介绍说明，请参见上述可选实施例对于第一序列样本的静态特征及其获取过程、第一序列样本的优化向量表示的计算过程的介绍说明，此处不再赘述。

步骤 24E，采用第二序列样本和第二序列样本的隐向量表示，对预训练的第一子模型和第三子模型进行联合训练，得到序列挖掘模型。

计算机设备可以采用第二序列样本和第二序列样本的隐向量表示，对预训练的第一子模型和第三子模型进行联合训练，即采用第二序列样本对预训练的第一子模型进行训练，同时采用第二序列样本的隐向量表示对第三子模型进行训练，从而得到序列挖掘模型，该序列挖掘模型包括联合训练的第一子模型和联合训练的第三子模型。

在一个示例中，上述步骤 24E 包括如下几个子步骤：计算第一损失函数的值和第三损失函数的值，第一损失函数是指第一子模型的损失函数，第三损失函数是指第三子模型的损失函数；根据第一损失函数的值、第三损失函数的值和权重系数设置，计算序列挖掘模型的损失函数的值，权重系数设置是指对第一损失函数的权重和第三损失函数的权重的设置；根据序列挖掘模型的损失函数的值，调整序列挖掘模型的参数，得到序列挖掘模型。有关第一损失函数的值的具体计算过程，请参见上述可选实施例，此处不再赘述。

示例性地，在第二序列样本中既包括存在正标签信息的序列样本，也包括存在负标签信息的序列样本，则第三子模型可以实现为一个标准分类模型，进而第三损失函数的值可以通过如下表达式计算得到：

$$\text{classifier_loss} = -[y * \log(p) + (1 - y) * \log(1 - p)]$$

其中, $classifier_loss$ 表示第三子模型的损失函数的值; y 表示第一序列样本具备预测的标签信息, p 表示第二序列样本具备预测的标签信息的概率。

根据第一损失函数的值和第三损失函数的值, 可以计算得到序列挖掘模型的损失函数的值如下所示:

$$loss_{\text{联合训练}} = r_{\text{classifier}} * classifier_loss + r_{\text{ae}} * ae_loss$$

其中, $loss_{\text{联合训练}}$ 表示序列挖掘模型的损失函数的值, 即预训练的第一子模型和第三子模型进行联合训练时损失函数的值; $r_{\text{classifier}}$ 和 r_{ae} 表示权重系数设置。

示例性地, 在第二序列样本中仅包括存在正标签信息的序列样本, 或者仅包括存在负标签信息的序列样本时, 目前已有的标准分类模型无法对第二序列样本的隐向量表示进行处理, 此时第三子模型可以实现为 PULearning 模型, 则第三损失函数的值可以用过如下表达式计算得到:

$$\tilde{R}_{pu}(g) = \pi_p \hat{R}_p^+(g) + \max\{0, \hat{R}_u^-(g) - \pi_p \hat{R}_p^-(g)\}$$

$$\text{其中, } \hat{R}_p^+(g) = \left(\frac{1}{n_p}\right) \sum_{i=1}^{n_p} l(g(x_i^p), 1), \hat{R}_u^-(g) = \left(\frac{1}{n_u}\right) \sum_{i=1}^{n_u} l(g(x_i^u), -1), \\ \hat{R}_p^-(g) = \left(\frac{1}{n_p}\right) \sum_{i=1}^{n_p} l(g(x_i^p), -1)。$$

其中, $\tilde{R}_{pu}$ 表示第三损失函数的值; g 为决策函数; x^p 表示存在正标签信息的序列样本或者存在负标签信息的序列样本, 即第二序列样本; x^u 表示不存在标签信息的序列样本, 即第三序列样本。

根据第一损失函数的值和第三损失函数的值, 可以计算得到序列挖掘模型的损失函数的值如下所示:

$$loss_{\text{联合训练}} = r_{pu} * \tilde{R}_{pu} + r_{\text{ae}} * ae_loss$$

其中, $loss_{\text{联合训练}}$ 表示序列挖掘模型的损失函数的值, 即预训练的第一子模型和第三子模型进行联合训练时损失函数的值; r_{pu} 和 r_{ae} 表示权重系数设置。

综上所述, 本申请实施例提供的技术方案, 在标签信息为序列样本至少部分存在标签信息, 且存在标签信息的序列样本的数量大于不存在标签信息的序列样本的数量时, 确定序列挖掘模型包括第一子模型和第三子模型。并且, 本申请实施例在对序列挖掘模型进行训练时, 首先对第一子模型进行预训练, 得到预训练的第一子模型, 由于预训练的第一子模型的损失函数收敛, 从而序列样本经过预训练的第一子模型得到的隐向量表示较为可靠; 接着, 采用存在标签信息的序列样本及其隐向量表示对预训练的第一子模型和第三子模型进行联合训练, 得到序列挖掘模型, 从而提供了一种在序列样本至少部分存在标签信息时, 序列挖掘模型的具体训练方法。

另外, 本申请实施例提供的技术方案中, 在至少部分存在标签信息的序列样本既包括存在正标签信息的序列样本, 也包括存在负标签信息的序列样本时, 提出了将第三子模型具体实现为标准分类模型; 在至少部分存在标签信息的序列样本仅包括存在正标签信息的序列样本, 或者仅包括存在负标签信息的序列样本时, 提出了将第三子模型具体实现为 PULearning 模型, 从而针对不同业务场景中序列数据不同的标签状态, 均可以构建序列挖掘模型, 避免

了某些历史序列数据无法使用，提升了数据资源的利用率，同时提升了序列挖掘模型的准确性。

在又一种可能的实施方式中，上述标签状态为第一序列样本至少部分存在标签信息，且第二序列样本的数量小于第三序列样本的数量；其中，第二序列样本是指第一序列样本中存在标签信息的序列样本；第三序列样本是指第一序列样本中不存在标签信息的序列样本；上述序列挖掘模型包括第一子模型、第二子模型和第三子模型；如图9所示，上述步骤240包括如下几个子步骤（24B~24J）：

步骤24B，采用第一序列样本对第一子模型进行预训练，得到预训练的第一子模型。

有关对第一子模型进行预训练过程的详细介绍说明，请参见上述可选实施例，此处不再赘述。需要说明的一点是，本申请实施例中也可以采用第二序列样本或者第三序列样本对第一子模型进行预训练，得到预训练的第一子模型，从而可以提升计算机设备对第一子模型进行预训练的速度，降低计算机设备的处理开销。

步骤24D，通过预训练的第一子模型对第二序列样本进行处理，得到第二序列样本的隐向量表示。

由于第一序列样本中不存在标签信息的第三序列样本的数量较多，此时如果直接采用存在标签信息的第二序列样本训练第三子模型，最终通过第三子模型确定序列数据的标签信息，可能会带来确定结果不准确的问题，因为存在标签信息的序列样本的数量较少，训练出的第三子模型不够准确。但此时，如果直接采用不存在标签信息的第三序列样本训练第二子模型，最终通过第二子模型确定序列数据的标签信息，这样忽略了存在标签信息的第二序列样本，造成数据资源浪费的同时，也使得第二子模型不够准确。基于此，本申请实施例提出了通过第一子模型、第二子模型和第三子模型构建序列挖掘模型，然后不同的序列样本分别对这些子模型进行训练，不仅充分利用了数据资源，也使得训练出的序列挖掘模型更加准确。

由于目前已有的可以实现为第三子模型的分类模型多用于处理二维的序列数据，因此，计算机设备在得到预训练的第一子模型后，可以通过该预训练的第一子模型对第二序列样本进行处理，以得到第二序列样本的隐向量表示，从而在预训练的第一子模型和第三子模型的联合训练过程中，采用第二序列样本中的关键特征信息对第三子模型进行训练，避免采用三维的第二序列样本对第三子模型进行训练，一方面可以利用目前已有的分类器作为第三子模型，另一方面可以降低计算机设备训练第三子模型时的处理开销。

可选地，为了提升序列挖掘模型的准确性和数据资源的利用率，计算机设备在得到第二序列样本的隐向量表示之后，还包括：获取第二序列样本的静态特征；根据第二序列样本的静态特征和第二序列样本的隐向量表示，得到第二序列样本的优化向量表示；其中，第二序列样本的优化向量表示用于对预训练的第一子模型和第三子模型进行联合训练。有关第二序列样本的静态特征及其获取过程、第二序列样本的优化向量表示的计算过程的详细介绍说明，请参见上述可选实施例对于第一序列样本的静态特征及其获取过程、第一序列样本的优化向量表示的计算过程的介绍说明，此处不再赘述。

步骤24F，采用第二序列样本和第二序列样本的隐向量表示，对预训练的第一子模型和第三子模型进行联合训练，得到联合训练的第一子模型。

计算机设备可以采用第二序列样本和第二序列样本的隐向量表示，对预训练的第一子模型和第三子模型进行联合训练，即采用第二序列样本对预训练的第一子模型进行训练，同时

采用第二序列样本的隐向量表示对第三子模型进行训练，从而得到序列挖掘模型，序列挖掘模型包括联合训练的第一子模型和联合训练的第三子模型。由于经过了联合训练，联合训练的第一子模型中引入了序列样本的标签信息，从而序列样本经过联合训练的第一子模型后得到的隐向量表示更为可靠。

步骤 24H，通过联合训练的第一子模型对第三序列样本进行处理，得到第三序列样本的隐向量表示。

在得到联合训练的第一子模型后，计算机设备可以通过该联合训练的第一子模型对第三序列样本进行处理，得到第三序列样本的隐向量表示。可选地，为了提升序列挖掘模型的准确性和数据资源的利用率，计算机设备在得到第三序列样本的隐向量表示之后，还包括：获取第三序列样本的静态特征；根据第三序列样本的静态特征和第三序列样本的隐向量表示，得到第三序列样本的优化向量表示；其中，第三序列样本的优化向量表示用于对联合训练的第一子模型和第二子模型进行联合训练。有关第三序列样本的静态特征及其获取过程、第三序列样本的优化向量表示的计算过程的详细介绍说明，请参见上述可选实施例对于第一序列样本的静态特征及其获取过程、第一序列样本的优化向量表示的计算过程的介绍说明，此处不再赘述。

步骤 24J，采用第三序列样本和第三序列样本的隐向量表示，对联合训练的第一子模型和第二子模型进行联合训练，得到序列挖掘模型。

计算机设备可以采用第三序列样本和第三序列样本的隐向量表示，对联合训练的第一子模型和第二子模型进行联合训练，即采用第三序列样本对联合训练的第一子模型进行训练，同时采用第三序列样本的隐向量表示对第二子模型进行训练，从而得到序列挖掘模型。

综上所述，本申请实施例提供的技术方案，在标签信息为序列样本至少部分存在标签信息，且存在标签信息的序列样本的数量小于不存在标签信息的序列样本的数量时，确定序列挖掘模型包括第一子模型、第二子模型和第三子模型。并且，本申请实施例在对序列挖掘模型进行训练时，首先对第一子模型进行预训练，得到预训练的第一子模型，由于预训练的第一子模型的损失函数收敛，从而序列样本经过预训练的第一子模型得到的隐向量表示较为可靠；接着，采用存在标签信息的序列样本及其隐向量表示对预训练的第一子模型和第三子模型进行联合训练，得到联合训练的第一子模型，由于联合训练的第一子模型中引入了序列样本的标签信息，从而序列样本经过联合训练的第一子模型得到的隐向量更为可靠；最后，采用不存在标签信息的序列样本及其隐向量表示对联合训练的第一子模型和第二子模型进行联合训练，得到序列挖掘模型，从而提供了一种在序列样本至少部分存在标签信息时，序列挖掘模型的具体训练方法。并且，本申请实施例提供的技术方案中，充分利用了目标业务场景下的历史序列数据，相比于相关技术中只能利用部分历史序列数据训练模型，本申请实施例提供的技术方案进一步提升了数据资源的利用率，并且使得训练出的序列挖掘模型更为准确。

请参考图 10，其示出了本申请一个实施例提供的序列数据的处理方法的流程图。该方法可应用于上文介绍的计算机设备中。该方法可以包括如下几个步骤（1010~1030）：

步骤 1010，获取目标业务场景的序列数据。

步骤 1020，调用目标业务场景对应的序列挖掘模型，序列挖掘模型是根据目标业务场景下的历史序列数据的标签状态，从序列挖掘框架中选取子模型构建并采用历史序列数据进行训练得到的模型。

步骤 1030, 通过序列挖掘模型对序列数据进行处理, 得到序列数据的标签信息。

其中, 序列挖掘框架包括第一子模型、第二子模型和第三子模型; 第一子模型用于获取序列数据的隐向量表示; 第二子模型用于在标签状态满足第一条件的情况下, 根据序列数据的隐向量表示确定序列数据的标签信息, 第一条件包括标签状态为第一序列样本不存在标签信息; 第三子模型用于在标签状态满足第二条件的情况下, 根据序列数据的隐向量表示确定序列数据的标签信息, 第二条件包括标签状态为第一序列样本至少部分存在标签信息。

有关目标业务场景、序列数据、序列挖掘框架、序列挖掘模型、标签信息、序列挖掘模型的训练过程等详细介绍说明, 请参见上述可选实施例, 此处不再赘述。

综上所述, 本申请实施例提供的技术方案, 通过获取目标业务场景的序列数据, 然后调用目标业务场景对应的序列挖掘模型, 对序列数据进行处理, 得到序列数据的标签信息, 提供了一种确定序列数据的标签信息的方式。并且, 本申请实施例提供的技术方案中, 序列挖掘模型是根据目标业务场景下的历史数据的标签状态, 从序列挖掘框架中选取对应的子模型构建而成的, 所以序列挖掘模型可以与目标业务场景相吻合, 提升对序列数据的标签信息确定的准确性。另外, 相比于现有技术只能针对具体业务场景开发单一的序列挖掘模型, 本申请实施例提供的序列挖掘框架可以针对不同的业务场景构建不同组成的序列挖掘模型, 具有较高的通用性。

请参考图 11, 其示出了本申请一个实施例提供的序列挖掘模型的训练方法和序列数据的处理方法的流程图。该方法可应用于上文介绍的计算机设备中。该方法可以包括如下几个步骤:

步骤 1110, 获取目标业务场景的历史序列数据, 并将该历史序列数据作为序列样本。目标业务场景是指需要确定其业务场景中序列数据的标签信息的业务场景, 序列数据, 又称为时间序列数据, 是指目标业务场景在不同时间上收集到的数据, 用于描述目标业务场景中出现的现象随时间变化的情况。

步骤 1120, 在目标业务场景中“好”与“坏”的长期行为模式上有明显不同的情况下, 分析序列样本的标签状态。序列样本的标签状态用于指示序列样本具备标签信息的情况, 标签信息用于反映第一序列样本的特征属性, 可选地, 标签信息可以包括“好”和“坏”两类, 其中, “好”是指与目标业务场景的主体业务相吻合, “坏”是指与目标业务场景的主体业务相背离。

步骤 1131, 在标签状态为序列样本不存在标签信息的情况下, 从序列挖掘框架中选取深度自编码模型和聚类模型构建序列挖掘模型。如图 12 所示, 其示出了本申请实施例提供的一种序列挖掘框架的示意图, 该序列挖掘框架包括深度自编码模型、聚类模型、标准分类模型和 PULearning 模型; 计算机设备从序列挖掘模型中选取深度自编码模型和聚类模型构建序列挖掘模型。

步骤 1132, 在标签状态为序列样本至少部分存在标签信息, 该标签信息既包括正标签信息, 也包括负标签信息, 且存在标签信息的序列样本的数量大于不存在标签信息的序列样本的数量时, 从序列挖掘框架中选取深度自编码模型和标准分类模型构建序列挖掘模型。正标签信息对应上文所述的“好”, 负标签信息对应上文所述的“坏”, 在存在标签信息的序列样本中既包括存在正标签信息的序列样本, 也包括存在负标签信息的序列样本时, 计算机设备从序列挖掘框架中选取的分类模型是标准分类模型。

步骤 1133, 在标签状态为序列样本至少部分存在标签信息, 该标签信息既包括正标签信息, 或仅包括负标签信息, 且存在标签信息的序列样本的数量大于不存在标签信息的序列样本的数量时, 从序列挖掘框架中选取深度自编码模型和 PULearning 模型构建序列挖掘模型; 在存在标签信息的序列样本中仅包括存在正标签信息的序列样本, 或者仅包括存在负标签信息的序列样本时, 计算机设备从序列挖掘框架中选取的分类模型是 PULearning 模型。

步骤 1134, 在标签状态为序列样本至少部分存在标签信息, 该标签信息既包括正标签信息, 也包括负标签信息, 且存在标签信息的序列样本的数量小于不存在标签信息的序列样本的数量时, 从序列挖掘框架中选取深度自编码模型、标准分类模型和聚类模型构建序列挖掘模型。由于存在标签信息的序列样本的数量小于不存在标签信息的序列样本的数量, 即序列样本中包含大量无标签的序列样本, 此时如果直接使用聚类模型对序列样本进行无监督学习, 就会忽略到存在标签信息的序列样本, 造成资源浪费, 因此本申请实施例提供的技术方案在序列挖掘模型中加入分类模型, 从而可以充分利用序列样本, 提升资源利用率, 提高序列挖掘模型的准确性。

步骤 1135, 在标签状态为序列样本至少部分存在标签信息, 该标签信息仅包括正标签信息, 或仅包括负标签信息, 且存在标签信息的序列样本的数量小于不存在标签信息的序列样本的数量时, 从序列挖掘框架中选取深度自编码模型、PULearning 模型和聚类模型构建序列挖掘模型。

步骤 1140, 采用序列样本对序列挖掘模型进行训练。序列挖掘模型可以用于确定目标业务场景下序列数据的标签信息, 计算机设备可以采用序列样本通过深度自编码模型后得到序列样本的隐向量表示, 然后采用序列样本及其隐向量表示对序列挖掘模型进行训练, 具体的训练过程参见上述可选实施例。

步骤 1150, 获取目标业务场景的序列数据。目标业务场景中不断有新的序列数据生成, 计算机设备可以获取该序列数据, 然后通过序列挖掘模型对序列数据进行处理。

步骤 1160, 通过序列挖掘模型对序列数据进行处理, 得到序列数据的隐向量表示。由于序列挖掘模型是根据目标业务场景的历史序列数据训练得到的, 且本申请实施例提供训练序列挖掘模型的方法充分利用了所有历史序列数据, 使得训练出的序列挖掘模型与目标业务场景高度适配, 从而通过该序列挖掘模型确定的序列数据的标签信息的准确性也较高。

请参考图 13, 其示出了本申请一个实施例提供的序列挖掘模型的训练装置的框图。该装置具有实现上述序列挖掘模型的训练方法示例的功能, 所述功能可以由硬件实现, 也可以由硬件执行相应的软件实现。该装置可以是计算机设备, 也可以设置在计算机设备中。该装置 1300 可以包括: 样本获取模块 1310、状态确定模块 1320、模型构建模块 1330 和模型训练模块 1340。

样本获取模块 1310, 用于获取目标业务场景下的第一序列样本, 所述第一序列样本包括所述目标业务场景下的历史序列数据。

状态确定模块 1320, 用于确定所述第一序列样本的标签状态, 所述第一序列样本的标签状态用于指示所述第一序列样本具备标签信息的情况。

模型构建模块 1330, 用于根据所述标签状态, 从序列挖掘框架中选取子模型构建序列挖掘模型, 所述序列挖掘模型用于确定所述目标业务场景下序列数据的标签信息。

模型训练模块 1340, 用于采用所述第一序列样本对所述序列挖掘模型进行训练。

其中, 所述序列挖掘框架包括第一子模型、第二子模型和第三子模型, 所述第一子模型用于获取所述序列数据的隐向量表示; 所述第二子模型用于在所述标签状态满足第一条件的情况下, 根据所述序列数据的隐向量表示确定所述序列数据的标签信息, 所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息; 所述第三子模型用于在所述标签状态满足第二条件的情况下, 根据所述序列数据的隐向量表示确定所述序列数据的标签信息, 所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

可选地, 所述标签状态为所述第一序列样本不存在标签信息, 所述序列挖掘模型包括所述第一子模型和所述第二子模型; 如图 14 所示, 所述模型训练模块 1340 包括: 预训练子模块 1342, 用于采用所述第一序列样本对所述第一子模型进行预训练, 得到预训练的第一子模型; 第一隐向量提取子模块 1344, 用于通过所述预训练的第一子模型对所述第一序列样本进行处理, 得到所述第一序列样本的隐向量表示; 第一模型训练子模块 1346, 用于采用所述第一序列样本和所述第一序列样本的隐向量表示, 对所述预训练的第一子模型和所述第二子模型进行联合训练, 得到所述序列挖掘模型。

可选地, 所述标签状态为所述第一序列样本至少部分存在标签信息, 且第二序列样本的数量大于第三序列样本的数量; 如图 14 所示, 其中, 所述第二序列样本是指所述第一序列样本中存在标签信息的序列样本; 所述第三序列样本是指所述第一序列样本中不存在标签信息的序列样本; 所述序列挖掘模型包括所述第一子模型和所述第三子模型; 所述模型训练模块 1340 包括: 预训练子模块 1342, 用于采用所述第一序列样本对所述第一子模型进行预训练, 得到预训练的第一子模型; 第二隐向量提取子模块 134A, 用于通过所述预训练的第一子模型对所述第二序列样本进行处理, 得到所述第二序列样本的隐向量表示; 第二模型训练子模块 134C, 用于采用所述第二序列样本和所述第二序列样本的隐向量表示, 对所述预训练的第一子模型和所述第三子模型进行联合训练, 得到所述序列挖掘模型。

可选地, 所述标签状态为所述第一序列样本至少部分存在标签信息, 且第二序列样本的数量小于第三序列样本的数量; 其中, 所述第二序列样本是指所述第一序列样本中存在标签信息的序列样本; 所述第三序列样本是指所述第一序列样本中不存在标签信息的序列样本; 所述序列挖掘模型包括所述第一子模型、所述第二子模型和所述第三子模型; 如图 14 所示, 所述模型训练模块 1340 包括: 预训练子模块 1342, 用于采用所述第一序列样本对所述第一子模型进行预训练, 得到预训练的第一子模型; 第三隐向量提取子模块 134B, 用于通过所述预训练的第一子模型对所述第二序列样本进行处理, 得到所述第二序列样本的隐向量表示; 第三模型训练子模块 134D, 用于采用所述第二序列样本和所述第二序列样本的隐向量表示, 对所述预训练的第一子模型和所述第三子模型进行联合训练, 得到联合训练的第一子模型; 第四隐向量提取子模块 134F, 用于通过所述联合训练的第一子模型对所述第三序列样本进行处理, 得到所述第三序列样本的隐向量表示; 第四模型训练子模块 134H, 用于采用所述第三序列样本和所述第三序列样本的隐向量表示, 对所述联合训练的第一子模型和所述第二子模型进行联合训练, 得到所述序列挖掘模型。

可选地, 所述第一子模型包括隐向量提取单元和重构单元; 如图 14 所示, 所述预训练子模块 1342 包括: 隐向量提取单元 13422, 用于通过所述隐向量提取单元对所述第一序列样本进行处理, 得到所述第一序列样本的隐向量表示; 数据重构单元 13424, 用于通过所述重构单元对所述第一序列样本的隐向量表示进行处理, 得到重构后的第一序列样本; 第一函数计算单元 13426, 用于根据所述第一序列样本和所述重构后的第一序列样本, 计算第一损失函

数的值, 所述第一损失函数是指所述第一子模型对应的损失函数; 第一模型训练单元 13420, 用于根据所述第一损失函数的值, 调整所述第一子模型的参数, 得到所述预训练的第一子模型。

可选地, 所述隐向量提取单元包括映射子单元、降维子单元和递归子单元; 如图 14 所示, 所述隐向量提取单元 13422 用于: 通过所述映射子单元获取所述第一序列样本的特征信息, 得到所述第一序列样本的特征信息表示; 通过所述降维子单元对所述第一序列样本的特征信息表示进行降维处理, 得到所述第一序列样本的降维特征信息; 通过所述递归子单元学习所述第一序列样本的降维特征信息的序列间跨步长的变化规律, 得到所述第一序列样本的隐向量表示。

可选地, 如图 14 所示, 所述装置 1300 还包括: 特征获取模块 1348, 用于获取所述第一序列样本的静态特征; 向量优化模块 1349, 用于根据所述第一序列样本的静态特征和所述第一序列样本的隐向量表示, 得到所述第一序列样本的优化向量表示; 其中, 所述第一序列样本的优化向量表示用于对所述预训练的第一子模型和所述第二子模型进行联合训练。

可选地, 如图 14 所示, 所述向量优化模块 1349 用于: 对所述第一序列样本的静态特征进行正则化处理, 得到正则化处理后的静态特征; 对所述第一序列样本的隐向量表示进行泛化处理, 得到泛化处理后的隐向量表示; 结合所述正则化处理后的静态特征和所述泛化处理后的隐向量表示, 得到所述第一序列样本的优化向量表示。

可选地, 如图 14 所示, 所述第一模型训练子模块 1346 包括: 第二函数计算单元 13462, 用于计算第一损失函数的值和第二损失函数的值, 所述第一损失函数是指所述第一子模型的损失函数, 所述第二损失函数是指所述第二子模型的损失函数; 第三函数计算单元 13464, 用于根据所述第一损失函数的值、所述第二损失函数的值和权重系数设置, 计算所述序列挖掘模型的损失函数的值, 所述权重系数设置是指对所述第一损失函数的权重和所述第二损失函数的权重的设置; 第二模型训练单元 13466, 用于根据所述序列挖掘模型的损失函数的值, 调整所述序列挖掘模型的参数, 得到所述序列挖掘模型。

可选地, 如图 14 所示, 所述第二函数计算单元 13462 用于: 确定所述第一序列样本的 k 个类别, 所述 k 为正整数; 根据所述 k 个类别, 通过所述第二子模型初始化 k 个聚类中心; 计算所述第一序列样本至所述 k 个聚类中心的距离, 得到样本概率分布, 所述样本概率分布用于指示所述第一序列样本分别属于所述 k 个类别的概率; 根据所述样本概率分布计算仿真样本概率分布, 所述仿真样本概率分布用于指示在所述第一序列样本存在标签信息的情况下, 所述第一序列样本分别属于所述 k 个类别的概率; 将所述样本概率分布和所述仿真样本概率分布的相对熵确定为所述第二损失函数的值。

综上所述, 本申请实施例提供的技术方案, 提出了一种序列挖掘框架, 实际应用中通过获取目标业务场景下的序列样本, 根据该序列样本的标签状态从序列挖掘框架中选取子模型构建序列挖掘模型, 然后采用该序列样本对序列挖掘模型进行训练, 从而使得训练出的序列挖掘模型可以用来判断目标业务场景下序列数据的标签信息。并且, 本申请实施例提出的序列挖掘框架可以在不同的业务场景中通用, 该序列挖掘框架中包括至少一个子模型, 针对不同的业务场景下历史序列数据不同的标签状态, 可以从序列挖掘框架中选取对应的子模型构建序列挖掘模型。另外, 相比于相关技术中只能采用结构固定的单一模型进行训练, 可能造成某一业务场景下部分历史序列数据无法使用, 如相关技术中只能采用无监督聚类模型进行训练, 可能造成某一业务场景下存在标签信息的历史序列数据无法使用, 产生资源浪费的问

题, 本申请实施例提供的技术方案, 针对历史序列数据不同的标签状态, 可以根据不同的历史序列数据, 对不同的序列挖掘模型进行训练, 从而充分利用了业务场景中的所有历史序列数据, 提升了资源利用率。此外, 由于本申请实施例提供的技术方案中, 可以根据目标业务场景下所有的历史序列数据充分训练序列挖掘模型, 从而提升了通过该序列挖掘模型确定序列数据的标签信息的准确性。

请参考图 15，其示出了本申请一个实施例提供的序列数据的处理装置的框图。该装置具有实现上述序列数据的处理方法示例的功能，所述功能可以由硬件实现，也可以由硬件执行相应的软件实现。该装置可以是计算机设备，也可以设置在计算机设备中。该装置 1500 可以包括：数据获取模块 1510、模型调用模块 1520 和标签确定模块 1530。

数据获取模块 1510，用于获取目标业务场景的序列数据。

模型调用模块 1520，用于调用所述目标业务场景对应的序列挖掘模型，所述序列挖掘模型是根据所述目标业务场景下的历史序列数据的标签状态，从序列挖掘框架中选取子模型构建并采用所述历史序列数据进行训练得到的模型。

标签确定模块 1530，用于通过所述序列挖掘模型对所述序列数据进行处理，得到所述序列数据的标签信息。

其中，所述序列挖掘框架包括第一子模型、第二子模型和第三子模型；所述第一子模型用于获取所述序列数据的隐向量表示；所述第二子模型用于在所述标签状态满足第一条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息；所述第三子模型用于在所述标签状态满足第二条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

综上所述，本申请实施例提供的技术方案，通过获取目标业务场景的序列数据，然后调用目标业务场景对应的序列挖掘模型，对序列数据进行处理，得到序列数据的标签信息，提供了一种确定序列数据的标签信息的方式。并且，本申请实施例提供的技术方案中，序列挖掘模型是根据目标业务场景下的历史数据的标签状态，从序列挖掘框架中选取对应的子模型构建而成的，所以序列挖掘模型可以与目标业务场景相吻合，提升对序列数据的标签信息确定的准确性。另外，相比于现有技术只能针对具体业务场景开发单一的序列挖掘模型，本申请实施例提供的序列挖掘框架可以针对不同的业务场景构建不同组成的序列挖掘模型，具有较高的通用性。

需要说明的是，上述实施例提供的装置，在实现其功能时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的装置与方法实施例属于同一构思，其具体实现过程详见方法实施例，这里不再赘述。

请参考图 16，其示出了本申请实施例提供的计算机设备的结构框图，该计算机设备可用于实现上述序列挖掘模型的训练方法示例，或者上述序列数据的处理方法示例的功能。具体来讲：

该计算机设备 1600 包括处理单元 (如 CPU (Central Processing Unit, 中央处理器)、GPU

(Graphics Processing Unit, 图形处理器) 和 FPGA (Field Programmable Gate Array, 现场可编程逻辑门阵列) 等) 1601、包括 RAM (Random-Access Memory, 随机存储器) 1602 和 ROM (Read-Only Memory, 只读存储器) 1603 的系统存储器 1604, 以及连接系统存储器 1604 和中央处理单元 1601 的系统总线 1605。该计算机设备 1600 还包括帮助计算计算机设备内的各个器件之间传输信息的 I/O 系统 (Input Output System, 基本输入/输出系统) 1606, 和用于存储操作系统 1613、应用程序 1614 和其他程序模块 1615 的大容量存储设备 1607。

该基本输入/输出系统 1606 包括有用于显示信息的显示器 1608 和用于用户输入信息的诸如鼠标、键盘之类的输入设备 1609。其中, 该显示器 1608 和输入设备 1609 都通过连接到系统总线 1605 的输入输出控制器 1610 连接到中央处理单元 1601。该基本输入/输出系统 1606 还可以包括输入输出控制器 1610 以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地, 输入输出控制器 1610 还提供输出到显示屏、打印机或其他类型的输出设备。

该大容量存储设备 1607 通过连接到系统总线 1605 的大容量存储控制器 (未示出) 连接到中央处理单元 1601。该大容量存储设备 1607 及其相关联的计算机可读介质为计算机设备 1600 提供非易失性存储。也就是说, 该大容量存储设备 1607 可以包括诸如硬盘或者 CD-ROM (Compact Disc Read-Only Memory, 只读光盘) 驱动器之类的计算机可读介质 (未示出)。

不失一般性, 该计算机可读介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括 RAM、ROM、EPROM (Erasable Programmable Read-Only Memory, 可擦写可编程只读存储器)、EEPROM (Electrically Erasable Programmable Read-Only Memory, 电可擦写可编程只读存储器)、闪存或其他固态存储其技术, CD-ROM、DVD (Digital Video Disc, 高密度数字视频光盘) 或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然, 本领域技术人员可知该计算机存储介质不局限于上述几种。上述的系统存储器 1604 和大容量存储设备 1607 可以统称为存储器。

根据本申请实施例, 该计算机设备 1600 还可以通过诸如因特网等网络连接到网络上的远程计算机运行。也即计算机设备 1600 可以通过连接在该系统总线 1605 上的网络接口单元 1611 连接到网络 1612, 或者说, 也可以使用网络接口单元 1611 来连接到其他类型的网络或远程计算机系统 (未示出)。

该存储器还包括至少一条指令、至少一段程序、代码集或指令集, 该至少一条指令、至少一段程序、代码集或指令集存储于存储器中, 且经配置以由一个或者一个以上处理器执行, 以实现上述序列挖掘模型的训练方法, 或者上述序列数据的处理方法。

在示例性实施例中, 还提供了一种计算机可读存储介质, 所述存储介质中存储有至少一条指令、至少一段程序、代码集或指令集, 所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现上述序列挖掘模型的训练方法, 或者上述序列数据的处理方法。

可选地, 该计算机可读存储介质可以包括: 只读存储器 (ROM, Read Only Memory)、随机存取记忆体 (RAM, Random Access Memory)、固态硬盘 (SSD, Solid State Drives) 或光盘等。其中, 随机存取记忆体可以包括电阻式随机存取记忆体 (ReRAM, Resistance Random

Access Memory) 和动态随机存取存储器 (DRAM, Dynamic Random Access Memory)。

在示例性实施例中, 还提供一种计算机程序产品, 当所述计算机程序产品在计算机设备上运行时, 使得计算机设备执行上述序列挖掘模型的训练方法, 或者上述序列数据的处理方法。

应当理解的是, 在本文中提及的“多个”是指两个或两个以上。“和/或”, 描述关联对象的关联关系, 表示可以存在三种关系, 例如, A 和/或 B, 可以表示: 单独存在 A, 同时存在 A 和 B, 单独存在 B 这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。另外, 本文中描述的步骤编号, 仅示例性示出了步骤间的一种可能的执行先后顺序, 在一些其它实施例中, 上述步骤也可以不按照编号顺序来执行, 如两个不同编号的步骤同时执行, 或者两个不同编号的步骤按照与图示相反的顺序执行, 本申请实施例对此不作限定。

以上所述仅为本申请的示例性实施例, 并不用以限制本申请, 凡在本申请的精神和原则之内, 所作的任何修改、等同替换、改进等, 均应包含在本申请的保护范围之内。

权利要求书

1、一种序列挖掘模型的训练方法，应用于计算机设备中，所述方法包括：

获取目标业务场景下的第一序列样本，所述第一序列样本包括所述目标业务场景下的历史序列数据；

确定所述第一序列样本的标签状态，所述第一序列样本的标签状态用于指示所述第一序列样本具备标签信息的情况；

根据所述标签状态，从序列挖掘框架中选取子模型构建序列挖掘模型，所述序列挖掘模型用于确定所述目标业务场景下序列数据的标签信息；

采用所述第一序列样本对所述序列挖掘模型进行训练；

其中，所述序列挖掘框架包括第一子模型、第二子模型和第三子模型；所述第一子模型用于获取所述序列数据的隐向量表示；所述第二子模型用于在所述标签状态满足第一条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息；所述第三子模型用于在所述标签状态满足第二条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

2、根据权利要求1所述的方法，其中，所述标签状态为所述第一序列样本不存在标签信息，所述序列挖掘模型包括所述第一子模型和所述第二子模型；

所述采用所述第一序列样本对所述序列挖掘模型进行训练，包括：

采用所述第一序列样本对所述第一子模型进行预训练，得到预训练的第一子模型；

通过所述预训练的第一子模型对所述第一序列样本进行处理，得到所述第一序列样本的隐向量表示；

采用所述第一序列样本和所述第一序列样本的隐向量表示，对所述预训练的第一子模型和所述第二子模型进行联合训练，得到所述序列挖掘模型。

3、根据权利要求1所述的方法，其中，所述标签状态为所述第一序列样本至少部分存在标签信息，且第二序列样本的数量大于第三序列样本的数量；其中，所述第二序列样本是指所述第一序列样本中存在标签信息的序列样本；所述第三序列样本是指所述第一序列样本中不存在标签信息的序列样本；所述序列挖掘模型包括所述第一子模型和所述第三子模型；

所述采用所述第一序列样本对所述序列挖掘模型进行训练，包括：

采用所述第一序列样本对所述第一子模型进行预训练，得到预训练的第一子模型；

通过所述预训练的第一子模型对所述第二序列样本进行处理，得到所述第二序列样本的隐向量表示；

采用所述第二序列样本和所述第二序列样本的隐向量表示，对所述预训练的第一子模型和所述第三子模型进行联合训练，得到所述序列挖掘模型。

4、根据权利要求1所述的方法，其中，所述标签状态为所述第一序列样本至少部分存在标签信息，且第二序列样本的数量小于第三序列样本的数量；其中，所述第二序列样本是指

所述第一序列样本中存在标签信息的序列样本；所述第三序列样本是指所述第一序列样本中不存在标签信息的序列样本；所述序列挖掘模型包括所述第一子模型、所述第二子模型和所述第三子模型；

所述采用所述第一序列样本对所述序列挖掘模型进行训练，包括：

采用所述第一序列样本对所述第一子模型进行预训练，得到预训练的第一子模型；

通过所述预训练的第一子模型对所述第二序列样本进行处理，得到所述第二序列样本的隐向量表示；

采用所述第二序列样本和所述第二序列样本的隐向量表示，对所述预训练的第一子模型和所述第三子模型进行联合训练，得到联合训练的第一子模型；

通过所述联合训练的第一子模型对所述第三序列样本进行处理，得到所述第三序列样本的隐向量表示；

采用所述第三序列样本和所述第三序列样本的隐向量表示，对所述联合训练的第一子模型和所述第二子模型进行联合训练，得到所述序列挖掘模型。

5、根据权利要求2至4任一项所述的方法，其中，所述第一子模型包括隐向量提取单元和重构单元；

所述采用所述第一序列样本对所述第一子模型进行预训练，得到预训练的第一子模型，包括：

通过所述隐向量提取单元对所述第一序列样本进行处理，得到所述第一序列样本的隐向量表示；

通过所述重构单元对所述第一序列样本的隐向量表示进行处理，得到重构后的第一序列样本；

根据所述第一序列样本和所述重构后的第一序列样本，计算第一损失函数的值，所述第一损失函数是指所述第一子模型对应的损失函数；

根据所述第一损失函数的值，调整所述第一子模型的参数，得到所述预训练的第一子模型。

6、根据权利要求5所述的方法，其中，所述隐向量提取单元包括映射子单元、降维子单元和递归子单元；

所述通过所述隐向量提取单元对所述第一序列样本进行处理，得到所述第一序列样本的隐向量表示，包括：

通过所述映射子单元获取所述第一序列样本的特征信息，得到所述第一序列样本的特征信息表示；

通过所述降维子单元对所述第一序列样本的特征信息进行降维处理，得到所述第一序列样本的降维特征信息；

通过所述递归子单元学习所述第一序列样本的降维特征信息的序列间跨步长的变化规律，得到所述第一序列样本的隐向量表示。

7、根据权利要求2所述的方法，其中，所述通过所述预训练的第一子模型对所述第一序

列样本进行处理,得到所述第一序列样本的隐向量表示之后,还包括:

获取所述第一序列样本的静态特征;

根据所述第一序列样本的静态特征和所述第一序列样本的隐向量表示,得到所述第一序列样本的优化向量表示;

其中,所述第一序列样本的优化向量表示用于对所述预训练的第一子模型和所述第二子模型进行联合训练。

8、根据权利要求7所述的方法,其中,所述根据所述第一序列样本的静态特征和所述第一序列样本的隐向量表示,得到所述第一序列样本的优化向量表示,包括:

对所述第一序列样本的静态特征进行正则化处理,得到正则化处理后的静态特征;

对所述第一序列样本的隐向量表示进行泛化处理,得到泛化处理后的隐向量表示;

结合所述正则化处理后的静态特征和所述泛化处理后的隐向量表示,得到所述第一序列样本的优化向量表示。

9、根据权利要求2所述的方法,其中,所述采用所述第一序列样本和所述第一序列样本的隐向量表示,对所述预训练的第一子模型和所述第二子模型进行联合训练,得到所述序列挖掘模型,包括:

计算第一损失函数的值和第二损失函数的值,所述第一损失函数是指所述第一子模型的损失函数,所述第二损失函数是指所述第二子模型的损失函数;

根据所述第一损失函数的值、所述第二损失函数的值和权重系数设置,计算所述序列挖掘模型的损失函数的值,所述权重系数设置是指对所述第一损失函数的权重和所述第二损失函数的权重的设置;

根据所述序列挖掘模型的损失函数的值,调整所述序列挖掘模型的参数,得到所述序列挖掘模型。

10、根据权利要求9所述的方法,其中,所述计算第二损失函数的值,包括:

确定所述第一序列样本的k个类别,所述k为正整数;

根据所述k个类别,通过所述第二子模型初始化k个聚类中心;

计算所述第一序列样本至所述k个聚类中心的距离,得到样本概率分布,所述样本概率分布用于指示所述第一序列样本分别属于所述k个类别的概率;

根据所述样本概率分布计算仿真样本概率分布,所述仿真样本概率分布用于指示在所述第一序列样本存在标签信息的情况下,所述第一序列样本分别属于所述k个类别的概率;

将所述样本概率分布和所述仿真样本概率分布的相对熵确定为所述第二损失函数的值。

11、一种序列数据的处理方法,应用于计算机设备中,所述方法包括:

获取目标业务场景的序列数据;

调用所述目标业务场景对应的序列挖掘模型,所述序列挖掘模型是根据所述目标业务场景下的历史序列数据的标签状态,从序列挖掘框架中选取子模型构建并采用所述历史序列数据进行训练得到的模型;

通过所述序列挖掘模型对所述序列数据进行处理，得到所述序列数据的标签信息；

其中，所述序列挖掘框架包括第一子模型、第二子模型和第三子模型；所述第一子模型用于获取所述序列数据的隐向量表示；所述第二子模型用于在所述标签状态满足第一条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息；所述第三子模型用于在所述标签状态满足第二条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

12、一种序列挖掘模型的训练装置，所述装置包括：

样本获取模块，用于获取目标业务场景下的第一序列样本，所述第一序列样本包括所述目标业务场景下的历史序列数据；

状态确定模块，用于确定所述第一序列样本的标签状态，所述第一序列样本的标签状态用于指示所述第一序列样本具备标签信息的情况；

模型构建模块，用于根据所述标签状态，从序列挖掘框架中选取子模型构建序列挖掘模型，所述序列挖掘模型用于确定所述目标业务场景下序列数据的标签信息；

模型训练模块，用于采用所述第一序列样本对所述序列挖掘模型进行训练；

其中，所述序列挖掘框架包括第一子模型、第二子模型和第三子模型；所述第一子模型用于获取所述序列数据的隐向量表示；所述第二子模型用于在所述标签状态满足第一条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息；所述第三子模型用于在所述标签状态满足第二条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

13、一种序列数据的处理装置，所述装置包括：

数据获取模块，用于获取目标业务场景的序列数据；

模型调用模块，用于调用所述目标业务场景对应的序列挖掘模型，所述序列挖掘模型是根据所述目标业务场景下的历史序列数据的标签状态，从序列挖掘框架中选取子模型构建并采用所述历史序列数据进行训练得到的模型；

标签确定模块，用于通过所述序列挖掘模型对所述序列数据进行处理，得到所述序列数据的标签信息；

其中，所述序列挖掘框架包括第一子模型、第二子模型和第三子模型；所述第一子模型用于获取所述序列数据的隐向量表示；所述第二子模型用于在所述标签状态满足第一条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第一条件包括所述标签状态为所述第一序列样本不存在标签信息；所述第三子模型用于在所述标签状态满足第二条件的情况下，根据所述序列数据的隐向量表示确定所述序列数据的标签信息，所述第二条件包括所述标签状态为所述第一序列样本至少部分存在标签信息。

14、一种计算机设备，所述计算机设备包括处理器和存储器，所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述

代码集或指令集由所述处理器加载并执行以实现如权利要求 1 至 10 任一项所述的序列挖掘模型的训练方法，或者实现如权利要求 11 所述的序列数据的处理方法。

15、一种计算机可读存储介质，所述存储介质中存储有至少一条指令、至少一段程序、代码集或指令集，所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现如权利要求 1 至 10 任一项所述的序列挖掘模型的训练方法，或者实现如权利要求 11 所述的序列数据的处理方法。