(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110569966 B
(45)授权公告日 2022.09.23

(21)申请号 201910848176.5
(22)申请日 2019.09.09
(65)同一申请的已公布的文献号
申请公布号 CN 110569966 A
(43)申请公布日 2019.12.13
(73)专利权人 联想(北京)有限公司
地址 100085 北京市海淀区上地信息产业基地创业路6号
(72)发明人 杨帆 金宝宝 张成松
(74)专利代理机构 北京集佳知识产权代理有限公司 11227
专利代理师 李金
(51)Int.Cl.
G06N 3/04(2006.01)
G06N 3/08(2006.01)
G06Q 10/06(2012.01)
G06Q 50/04(2012.01)

(56)对比文件
TW 200540674 A,2005.12.16
CN 109636014 A,2019.04.16
CN 106249724 A,2016.12.21

审查员 李卿

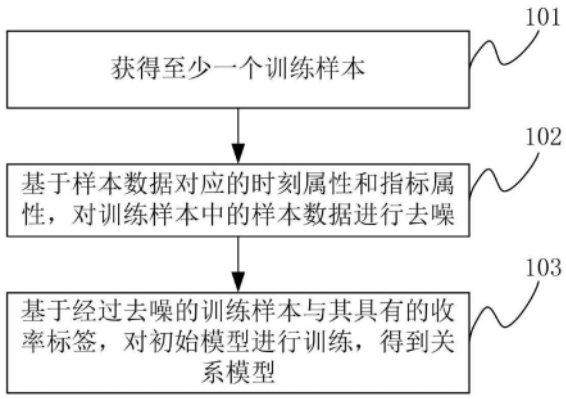
权利要求书2页 说明书11页 附图5页

(54)发明名称

一种数据处理方法、装置及电子设备

(57)摘要

本申请公开了一种数据处理方法、装置及电子设备，方法包括：获得至少一个训练样本，所述训练样本包括至少一个指标的指标数据，所述指标数据包括其所属指标在至少一个时刻下的样本数据，所述训练样本具有预设的收率标签；基于所述样本数据对应的时刻属性和指标属性，对所述训练样本中的样本数据进行去噪；基于经过去噪的训练样本与其具有的收率标签，对初始模型进行训练，得到关系模型，所述关系模型表征所述指标与收率之间的关系。可见，本申请中利用样本数据中的时刻属性和指标属性对样本数据进行去噪，使得用于训练关系模型的样本数据更加准确，由此使得利用这些样本数据所训练出的关系模型更加准确，从而达到提高关系模型准确率的目的。



CN 110569966 B

1.一种数据处理方法,包括:

获得至少一个训练样本,所述训练样本包括至少一个指标的指标数据,所述指标数据包括其所属指标在至少一个时刻下的样本数据,所述训练样本具有预设的收率标签;

利用自注意力机制,基于所述样本数据对应的时刻属性和指标属性,对所述训练样本中的样本数据进行去噪;

基于经过去噪的训练样本与其具有的收率标签,对初始模型进行训练,得到关系模型,所述关系模型表征所述指标与收率之间的关系。

2.根据权利要求1所述的方法,基于所述样本数据对应的时刻属性和指标属性,对所述训练样本中的样本数据进行去噪,包括:

获得目标指标的指标数据中的所述样本数据之间的关联关系;所述目标指标为所述至少一个指标中的任意一个;

基于所述关联关系,对所述目标指标的指标数据中的样本数据进行计算,以得到所述目标指标在所述时刻下的新的样本数据。

3.根据权利要求2所述的方法,所述关联关系包括所述目标指标的指标数据中的所述样本数据之间的相似度值;

其中,基于所述关联关系,对所述目标指标的指标数据中的样本数据进行计算,以得到所述目标指标在所述时刻下的新的样本数据,包括:

基于所述相似度值,确定所述目标指标的指标数据中在所述至少一个时刻中所述时刻对应的样本权重系数;

基于所述样本权重系数,对所述目标指标的指标数据中的样本数据进行计算,以得到所述目标指标在所述时刻下的新的样本数据。

4.根据权利要求1所述的方法,基于所述样本数据对应的时刻属性和指标属性,对所述训练样本中的样本数据进行去噪,包括:

基于所述样本数据对应的时刻属性和指标属性之间的关联关系,对所述样本数据进行更新,以使得所述训练样本中的样本数据实现去噪。

5.根据权利要求4所述的方法,利用所述样本数据对应的时刻属性和指标属性之间的关联关系,对所述样本数据进行更新,包括:

利用自注意力机制,基于所述样本数据对应的时刻属性和指标属性之间的关联关系,对所述样本数据进行更新。

6.根据权利要求1、2或4所述的方法,所述初始模型为基于预设的机器学习算法所搭建的拟合模型,所述初始模型具有至少一个与指标和收率相关的初始拟合参数。

7.一种数据处理装置,包括:

获得单元,用于获得至少一个训练样本,所述训练样本包括至少一个指标的指标数据,所述指标数据包括其所属指标在至少一个时刻下的样本数据,所述训练样本具有预设的收率标签;

去噪单元,用于利用自注意力机制,基于所述样本数据对应的时刻属性和指标属性,对所述训练样本中的样本数据进行去噪;

训练单元,用于基于经过去噪的训练样本与其具有的收率标签,对初始模型进行训练,得到关系模型,所述关系模型表征所述指标与收率之间的关系。

8.根据权利要求7所述的装置,所述去噪单元,包括:

关联获得子单元,用于获得目标指标的指标数据中的所述样本数据之间的关联关系;所述目标指标为所述至少一个指标中的任意一个;

样本计算子单元,用于基于所述关联关系,对所述目标指标的指标数据中的样本数据进行计算,以得到所述目标指标在所述时刻下的新的样本数据。

9.根据权利要求7或8所述的装置,还包括:

模型搭建单元,用于预先基于预设的机器学习算法搭建所述初始模型,所述初始模型为拟合模型,且所述初始模型具有至少一个与指标和收率相关的初始拟合参数。

10.一种电子设备,包括:

存储器,用于存储应用程序及应用程序运行所产生的数据;

处理器,用于执行所述应用程序,以实现:

获得至少一个训练样本,所述训练样本包括至少一个指标的指标数据,所述指标数据包括其所属指标在至少一个时刻下的样本数据,所述训练样本具有预设的收率标签;

利用自注意力机制,基于所述样本数据对应的时刻属性和指标属性,对所述训练样本中的样本数据进行去噪;

基于经过去噪的训练样本与其具有的收率标签,对初始模型进行训练,得到关系模型,所述关系模型表征所述指标与收率之间的关系。

一种数据处理方法、装置及电子设备

技术领域

[0001] 本申请涉及模型训练技术领域，尤其涉及一种数据处理方法、装置及电子设备。

背景技术

[0002] 对于制造行业中，设备的工艺指标直接决定了产品的收率。为了进行产品收率最大化，需要对各工艺指标进行寻优，即找到能够使得产品收率最大化的工艺指标。为此，目前通过机器学习训练工艺指标与产品收率之间的关系模型，进而在模型上获得能够让工艺指标最大化的工艺指标。

[0003] 但是制造行业中的工艺环境通常比较复杂，使得采集到的工艺指标的指标数据存在大量噪声，导致所训练出的模型准确性较低。

发明内容

[0004] 有鉴于此，本申请提供一种数据处理方法、装置及电子设备，用以提高对关系模型进行训练的准确率。

[0005] 本申请提供了一种数据处理方法，包括：

[0006] 获得至少一个训练样本，所述训练样本包括至少一个指标的指标数据，所述指标数据包括其所属指标在至少一个时刻下的样本数据，所述训练样本具有预设的收率标签；

[0007] 基于所述样本数据对应的时刻属性和指标属性，对所述训练样本中的样本数据进行去噪；

[0008] 基于经过去噪的训练样本与其具有的收率标签，对初始模型进行训练，得到关系模型，所述关系模型表征所述指标与收率之间的关系。

[0009] 上述方法，可选的，基于所述样本数据对应的时刻属性和指标属性，对所述训练样本中的样本数据进行去噪，包括：

[0010] 获得目标指标的指标数据中的所述样本数据之间的关联关系；所述目标指标为所述至少一个指标中的任意一个；

[0011] 基于所述关联关系，对所述目标指标的指标数据中的样本数据进行计算，以得到所述目标指标在所述时刻下的新的样本数据。

[0012] 上述方法，可选的，所述关联关系包括所述目标指标的指标数据中的所述样本数据之间的相似度值；

[0013] 其中，基于所述关联关系，对所述目标指标的指标数据中的样本数据进行计算，以得到所述目标指标在所述时刻下的新的样本数据，包括：

[0014] 基于所述相似度值，确定所述目标指标的指标数据中在所述至少一个时刻中所述时刻对应的样本权重系数；

[0015] 基于所述样本权重系数，对所述目标指标的指标数据中的样本数据进行计算，以得到所述目标指标在所述时刻下的新的样本数据。

[0016] 上述方法，可选的，基于所述样本数据对应的时刻属性和指标属性，对所述训练样

本中的样本数据进行去噪，包括：

[0017] 基于所述样本数据对应的时刻属性和指标属性之间的关联关系，对所述样本数据进行更新，以使得所述训练样本中的样本数据实现去噪。

[0018] 上述方法，可选的，利用所述样本数据对应的时刻属性和指标属性之间的关联关系，对所述样本数据进行更新，包括：

[0019] 利用自注意力机制，基于所述样本数据对应的时刻属性和指标属性之间的关联关系，对所述样本数据进行更新。

[0020] 上述方法，可选的，所述初始模型为基于预设的机器学习算法所搭建的拟合模型，所述初始模型具有至少一个与指标和收率相关的初始拟合参数。

[0021] 本申请还提供了一种数据处理装置，包括：

[0022] 获得单元，用于获得至少一个训练样本，所述训练样本包括至少一个指标的指标数据，所述指标数据包括其所属指标在至少一个时刻下的样本数据，所述训练样本具有预设的收率标签；

[0023] 去噪单元，用于基于所述样本数据对应的时刻属性和指标属性，对所述训练样本中的样本数据进行去噪；

[0024] 训练单元，用于基于经过去噪的训练样本与其具有的收率标签，对初始模型进行训练，得到关系模型，所述关系模型表征所述指标与收率之间的关系。

[0025] 上述装置，可选的，所述去噪单元，包括：

[0026] 关联获得子单元，用于获得目标指标的指标数据中的所述样本数据之间的关联关系；所述目标指标为所述至少一个指标中的任意一个；

[0027] 样本计算子单元，用于基于所述关联关系，对所述目标指标的指标数据中的样本数据进行计算，以得到所述目标指标在所述时刻下的新的样本数据。

[0028] 上述装置，可选的，还包括：

[0029] 模型搭建单元，用于预先基于预设的机器学习算法搭建所述初始模型，所述初始模型为拟合模型，且所述初始模型具有至少一个与指标和收率相关的初始拟合参数。

[0030] 本申请还提供了一种电子设备，包括：

[0031] 存储器，用于存储应用程序及应用程序运行所产生的数据；

[0032] 处理器，用于执行所述应用程序，以实现：

[0033] 获得至少一个训练样本，所述训练样本包括至少一个指标的指标数据，所述指标数据包括其所属指标在至少一个时刻下的样本数据，所述训练样本具有预设的收率标签；

[0034] 基于所述样本数据对应的时刻属性和指标属性，对所述训练样本中的样本数据进行去噪；

[0035] 基于经过去噪的训练样本与其具有的收率标签，对初始模型进行训练，得到关系模型，所述关系模型表征所述指标与收率之间的关系。

[0036] 从上述技术方案可以看出，本申请公开的一种数据处理方法、装置及电子设备，在获得到各指标在各时刻下的样本数据之后，基于这些样本数据中的时刻属性和指标属性，对样本数据进行去噪，再利用去燥后的训练样本与其具有的收率标签，对初始模型进行训练，以得到指标与收率之间的关系模型。可见，本申请中利用样本数据中的时刻属性和指标属性对样本数据进行去噪，使得用于训练关系模型的样本数据更加准确，由此使得利用这

些样本数据所训练出的关系模型更加准确，从而达到提高关系模型准确率的目的。

附图说明

[0037] 为了更清楚地说明本申请实施例的技术方案，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。

[0038] 图1为本申请实施例一提供的一种数据处理方法的流程图；

[0039] 图2及图3分别为本申请实施例的示例图；

[0040] 图4为本申请实施例二提供的一种数据处理装置的结构示意图；

[0041] 图5为本申请实施例二的部分结构示意图；

[0042] 图6为本申请实施例二的另一结构示意图；

[0043] 图7为本申请实施例三提供的电子设备的结构示意图；

[0044] 图8及图9分别为本申请实施例的应用示例图。

具体实施方式

[0045] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0046] 参考图1，为本申请实施例一提供的一种数据处理方法的流程图，该方法适用于具有数据处理能力并能够进行模型训练的电子设备中，如计算机或服务器等终端设备。本申请中的方法主要用于对用于训练指标与收率之间的关系模型的样本数据进行去噪，从而达到能够提高关系模型准确率的目的。

[0047] 具体的，本实施例中的方法可以包括以下步骤：

[0048] 步骤101：获得至少一个训练样本。

[0049] 其中，训练样本中包括至少一个指标的指标数据。而指标数据中包括其所属指标在至少一个时刻下的样本数据，也就是说，训练样本可以是由至少一个样本数据组成，当然还可以包括其他样本数据，而这些样本数据为在至少一个指标上的至少一个时刻下的样本数据，或者可以表述为：样本数据为至少一个时刻下的在至少一个指标上的样本数据。

[0050] 需要说明的是，这里的指标可以为制造行业中针对某个产品的产品生产指标，如温度指标、流量指标或压力指标等，每个指标在不同时刻下可能会有不同的采样数据。

[0051] 如图2中所示，训练样本中由3种指标的指标数据组成：y1、y2、y3，每种指标下的指标数据由该指标下在3个时刻上的样本数据组成：指标1在3个时刻上的样本数据：y11、y12、y13，指标2在3个时刻上的样本数据：y21、y22、y23，指标3在3个时刻上的样本数据：y31、y32、y33。当然训练样本还可以采用另一种表述：训练样本由3个时刻的时刻数据组成，每个时刻下的时刻数据由该时刻下在3个指标上的样本数据组成：时刻1在3个指标上的样本数据：y11、y21、y31，时刻2在3个指标上的样本数据：y12、y22、y32，时刻3在3个指标上的样本数据：y13、y23、y33。

[0052] 具体的，本实施例中在获得训练样本时，可以针对多个指标分别获得多个时刻下的数据作为样本数据，并以此组成一个(组)训练样本，或者说，可以在多个时刻下获得多个指标上的数据作为样本数据，并以此组成一个(组)训练样本。本实施例中可以采用这种方式获得一个或多个(组)训练样本。

[0053] 其中，训练样本中具有预设的收率标签，其中收率标签表明在样本数据下对应的实际收率值。具体的，本实施例中可以通过计算样本数据下的实际收率值，将计算得到的实际收率值来作为收率标签，或者，本实施例中可以针对样本数据在历史记录中采集对应的实际收率值进行采集，作为收率标签，等等。

[0054] 步骤102：基于样本数据对应的时刻属性和指标属性，对训练样本中的样本数据进行去噪。

[0055] 其中，样本数据中同一指标属性下的样本数据可能会因为时刻属性的不同而存在差异，因此，本实施例中对训练样本进行去噪，是基于样本数据对应的时刻属性和指标属性实现的，具体可以理解为：本实施例中在样本数据的指标属性的基础上，结合样本数据在时刻属性上的差异，对样本数据进行去噪；或者可以理解为，本实施例中在样本数据的时刻属性的基础上，结合样本数据在指标属性之间的关联，对样本数据进行去噪。

[0056] 例如，本实施例中分别在时刻属性和指标属性两种维度上，对样本数据中存在噪声的样本数据进行去噪；或者，本实施例中可以结合时刻属性和指标属性两种维度之间在样本数据上的互相影响，对样本数据中存在噪声的样本数据进行去噪；或者，本实施例中可以结合时刻属性和指标属性之间的关联关系，对样本数据中存在噪声的样本数据进行去噪，等等。

[0057] 步骤103：基于经过去噪的训练样本与其具有的收率标签，对初始模型进行训练，得到关系模型。

[0058] 其中，关系模型表征指标与收率之间的对应关系。

[0059] 具体的，本实施例中利用经过去噪的训练样本及其收率标签对初始模型进行训练，主要是对初始模型中的各种模型参数进行不断优化，直到模型参数最优，结束训练，此时初始模型基于优化的模型参数形成训练完成的关系模型。例如，本实施例中通过迭代的方式利用训练样本和收率标签对初始模型的模型参数进行优化，在每次的迭代训练中，监测模型参数是否收敛，即是否趋于稳定，随着多次训练，直到模型参数稳定，即前后两次训练的参数结果不再变化，此时确定为最优的模型参数，并由此得到训练完成的关系模型，该关系模型可以用于对产品收率最大化的指标进行寻优。

[0060] 由以上方案可知，本申请实施例一提供的一种数据处理方法，在获得到各指标在各时刻下的样本数据之后，基于这些样本数据中的时刻属性和指标属性，对样本数据进行去噪，再利用去燥后的训练样本与其具有的收率标签，对初始模型进行训练，以得到指标与收率之间的关系模型。可见，本实施例中利用样本数据中的时刻属性和指标属性对样本数据进行去噪，使得用于训练关系模型的样本数据更加准确，由此使得利用这些样本数据所训练出的关系模型更加准确，从而达到提高关系模型准确率的目的。

[0061] 在一种实现方式中，本实施例中的步骤102在对样本数据进行去噪时，具体可以通过以下方式实现：

[0062] 首先，获得目标指标的指标数据中的样本数据之间的关联关系，再基于关联关系，

对目标指标的指标数据中的样本数据分别进行计算，以得到目标指标在时刻下的新的样本数据。

[0063] 其中，目标指标为训练样本中至少一个指标中的任意一个，也就是说，本实施例中对于训练样本中的任意一个指标均执行以下操作：

[0064] 获得目标指标在至少一个时刻中的每个时刻下的样本数据，并获得这些样本数据之间的关联关系，即获得在目标指标上每个时刻下的样本数据之间的关联关系，再基于关联关系，对目标指标上每个时刻下的样本数据进行重新计算，得到目标指标上每个时刻下的新的样本数据。

[0065] 举例如下：对于训练样本中由3种指标在3个时刻下的样本数据：y11、y12、y13、y21、y22、y23、y31、y32、y33，针对每个指标在3个时刻下的样本数据分别执行以下操作：

[0066] 对指标1在3个时刻下的样本数据y11、y12、y13，分别计算y11、y12、y13之间的关联关系，然后基于y11、y12、y13之间的关联关系，对y11、y12、y13分别进行重新计算，得到指标1在3个时刻下的新的样本数据y11`、y12`、y13`；

[0067] 对指标2在3个时刻下的样本数据y21、y22、y23，分别计算y21、y22、y23之间的关联关系，然后基于y21、y22、y23之间的关联关系，对y21、y22、y23分别进行重新计算，得到指标2在3个时刻下的新的样本数据y21`、y22`、y23`；

[0068] 对指标3在3个时刻下的样本数据y31、y32、y33，分别计算y31、y32、y33之间的关联关系，然后基于y31、y32、y33之间的关联关系，对y31、y32、y33分别进行重新计算，得到指标3在3个时刻下的新的样本数据y31`、y32`、y33`。

[0069] 具体的，本实施例中的样本数据之间的关联关系可以为目标指标的指标数据中，各样本数据之间的相似度值。如y31、y32、y33之间的相似度值等。

[0070] 相应的，在基于关联关系对样本数据进行计算，得到新的样本数据时，可以通过以下方式实现：

[0071] 首先，基于相似度值，确定目标指标的指标数据中在至少一个时刻中的时刻对应的样本权重系数，即基于相似度值，确定目标指标上在各个时刻下的样本数据的样本权重系数，如对于y11来说，其样本权重系数包括：针对y11、y12、y13分别有：0.1、0.3、0.6；对于y12来说，其样本权重系数包括：针对y11、y12、y13分别有：0.2、0.4、0.4，对于y13来说，其样本权重系数包括：针对y11、y12、y13分别有：0.3、0.4、0.3；

[0072] 之后，基于样本权重系数，对目标指标的指标数据中的样本数据进行计算，以得到目标指标在时刻下的新的样本数据，也就是说，利用样本权重系数，对目标指标的指标数据中的每个样本数据分别进行重新计算，以得到每个时刻下新的样本数据。

[0073] 例如，在得到每个时刻下的样本数据的样本权重系数之后，利用该样本数据的样本权重系数结合与该样本数据所属同一目标指标的其他样本数据进行计算，得到新的样本数据。

[0074] 举例如下：如对于y11来说，y11的样本权重系数包括：针对y11、y12、y13分别有：0.1、0.3、0.6，将0.1*y11+0.3*y12+0.6*y13得到的值作为新的y11`；同理，对于y12来说，y12样本权重系数包括：针对y11、y12、y13分别有：0.2、0.4、0.4，将0.2*y11+0.4*y12+0.4*y13得到的值作为新的y12`；对于y13来说，y13样本权重系数包括：针对y11、y12、y13分别有：0.3、0.4、0.3，将0.3*y11+0.4*y12+0.3*y13得到的值作为新的y13`，如图3中所示。

[0075] 可见,本实施例中所得到的新的样本数据首先综合了多个样本数据计算了多个样本数据之间的相关性如相似度值,然后利用多个样本数据之间的相似度值,重新生成了每个样本数据,由此,有效利用了指标之间以及样本之间的数据,可以有效的减少噪声数据,提高所训练出的关系模型的准确性。

[0076] 需要说明的是,本实施例中实现以上样本数据的去噪可以通过自注意力机制(self-attention)实现。

[0077] 在另一种实现方式中,本实施例中的步骤102在对样本数据进行去噪时,具体可以通过以下方式实现:

[0078] 基于样本数据对应的时刻属性和指标属性之间的关联关系,对样本数据进行更新,以使得训练样本中的样本数据实现去噪。

[0079] 其中,样本数据在时刻属性和指标属性之间具有一定的关联关系,例如,在同一个指标属性上的不同时刻属性对应的样本数据之间具有相似性或者呈现具有一定规律的变化关系,如递增或递件或在一定幅值范围内不断变化等等;再如,在同一时刻属性对应的不同指标属性上的样本数据之间具有一定的对应关系,如此消彼长、线性递增或递减、非线性或者指数等关系,由此,本实施例中,结合样本数据对应的时刻属性和指标属性之间的关联关系,对样本数据进行更新,例如,对样本数据进行修改(增加或降低等)或重新生成新的样本数据,从而实现对样本数据的去噪处理。

[0080] 具体实现中,本实施例在进行去噪时,可以采用预设的去噪机制,例如,利用自注意力机制,基于样本数据对应的时刻属性和指标属性之间的关联关系,对样本数据进行更新。

[0081] 例如,利用自注意力机制,对样本数据在每个指标上的时刻维度上进行扩展,具体根据样本数据对应的时刻属性和指标属性之间的关联关系,扩展出在每个指标上的更多时刻上的样本数据,从而以数据扩展的方式,对样本数据进行更新,使得样本数据更加丰富,由此在大数据的层面使得具有噪声的样本数据所占据的比例较小,由此实现样本数据的去噪;

[0082] 或者,利用自注意力机制,对样本数据在每个指标上的时刻维度上进行修改,具体根据样本数据对应的时刻属性和指标属性之间的关联关系,对出在每个指标上的每个时刻上的样本数据进行修改,如按照比例增加或降低,或者直接在数值上进行增加或降低等,从而以数据修改的方式,对样本数据进行更新,使得样本数据更加准确,由此实现样本数据的去噪,等等。

[0083] 举例如下:对于训练样本中由3种指标在3个时刻下的样本数据:y11、y12、y13、y21、y22、y23、y31、y32、y33,结合在时刻属性和指标属性之间样本数据的关联关系,对样本数据生成新的表示,即新的样本数据,对样本数据可以在每个指标属性上扩展到m个维度(时刻)上的样本数据,此时,训练样本在3个指标属性上的每个指标上均具有m个样本数据,相应的,再基于这些扩展的样本数据进行模型训练,实现模型参数的优化,得到表征指标和收率之间的关系的关系模型,能够用于对收率最大化的指标进行寻优。

[0084] 可见,本实施例中所得到的新的样本数据综合了指标和时刻两种维度之间的关联,重新生成样本数据,由此,有效利用指标之间以及样本之间的关联,可以有效的减少噪声数据,提高所训练出的关系模型的准确性。

[0085] 基于以上实现，本实施例中的初始模型可以为基于预设的机器学习算法所搭建的拟合模型，该拟合模型具有至少一个初始拟合参数，本实施例中利用去噪后的训练样本及其收率标签对这些初始拟合参数进行训练优化，从而使得训练出的关系拟合模型更加准确，以便于更加准确的在收率最大化时对指标进行寻优。

[0086] 其中，机器学习算法可以为神经网络学习算法或卷积网络算法等。

[0087] 参考图4，为本申请实施例二提供的一种数据处理装置的结构示意图，该装置可以设置在具有数据处理能力并能够进行模型训练的电子设备中，如计算机或服务器等终端设备。本申请中的装置主要用于对用于训练指标与收率之间的关系模型的样本数据进行去噪，从而达到能够提高关系模型准确率的目的。

[0088] 具体的，本实施例中的装置可以包括以下功能单元：

[0089] 获得单元401，用于获得至少一个训练样本。

[0090] 其中，训练样本中包括至少一个指标的指标数据。而指标数据中包括其所属指标在至少一个时刻下的样本数据，也就是说，训练样本是由至少一个样本数据组成，而这些样本数据为在至少一个指标上的至少一个时刻下的样本数据，或者可以表述为：样本数据为至少一个时刻下的在至少一个指标上的样本数据。

[0091] 需要说明的是，这里的指标可以为制造行业中针对某个产品的产品生产指标，如温度指标、流量指标或压力指标等，每个指标在不同时刻下可能会有不同的采样数据。

[0092] 如图2中所示，训练样本中由3种指标的指标数据组成：y1、y2、y3，每种指标下的指标数据由该指标下在3个时刻上的样本数据组成：指标1在3个时刻上的样本数据：y11、y12、y13，指标2在3个时刻上的样本数据：y21、y22、y23，指标3在3个时刻上的样本数据：y31、y32、y33。当然训练样本还可以采用另一种表述：训练样本由3个时刻的时刻数据组成，每个时刻下的时刻数据由该时刻下在3个指标上的样本数据组成：时刻1在3个指标上的样本数据：y11、y21、y31，时刻2在3个指标上的样本数据：y12、y22、y32，时刻3在3个指标上的样本数据：y13、y23、y33。

[0093] 具体的，本实施例中在获得训练样本时，可以针对多个指标分别获得多个时刻下的数据作为样本数据，并以此组成一个(组)训练样本，或者说，可以在多个时刻下获得多个指标上的数据作为样本数据，并以此组成一个(组)训练样本。本实施例中可以采用这种方式获得一个或多个(组)训练样本。

[0094] 其中，训练样本中具有预设的收率标签，其中收率标签表明在样本数据下对应的实际收率值。具体的，本实施例中可以通过计算样本数据下的实际收率值，将计算得到的实际收率值来作为收率标签，或者，本实施例中可以针对样本数据在历史记录中采集对应的实际收率值进行采集，作为收率标签，等等。

[0095] 去噪单元402，用于基于所述样本数据对应的时刻属性和指标属性，对所述训练样本中的样本数据进行去噪；

[0096] 其中，样本数据中同一指标属性下的样本数据可能会因为时刻属性的不同而存在差异，因此，本实施例中对训练样本进行去噪，是基于样本数据对应的时刻属性和指标属性实现的，具体可以理解为：本实施例中在样本数据的指标属性的基础上，结合样本数据在时刻属性上的差异，对样本数据进行去噪；或者可以理解为，本实施例中在样本数据的时刻属性的基础上，结合样本数据在指标属性之间的关联，对样本数据进行去噪。

[0097] 例如，本实施例中分别在时刻属性和指标属性两种维度上，对样本数据中存在噪声的样本数据进行去噪；或者，本实施例中可以结合时刻属性和指标属性两种维度之间在样本数据上的互相影响，对样本数据中存在噪声的样本数据进行去噪；或者，本实施例中可以结合时刻属性和指标属性之间的关联关系，对样本数据中存在噪声的样本数据进行去噪，等等。

[0098] 训练单元403，用于基于经过去噪的训练样本与其具有的收率标签，对初始模型进行训练，得到关系模型。

[0099] 其中，关系模型表征指标与收率之间的对应关系。

[0100] 具体的，本实施例中利用经过去噪的训练样本及其收率标签对初始模型进行训练，主要是对初始模型中的各种模型参数进行不断优化，直到模型参数最优，结束训练，此时初始模型基于优化的模型参数形成训练完成的关系模型。例如，本实施例中通过迭代的方式利用训练样本和收率标签对初始模型的模型参数进行优化，在每次的迭代训练中，监测模型参数是否收敛，即是否趋于稳定，随着多次训练，直到模型参数稳定，即前后两次训练的参数结果不再变化，此时确定为最优的模型参数，并由此得到训练完成的关系模型，该关系模型可以用于对产品收率最大化的指标进行寻优。

[0101] 由以上方案可知，本申请实施例二提供的一种数据处理装置，在获得到各指标在各时刻下的样本数据之后，基于这些样本数据中的时刻属性和指标属性，对样本数据进行去噪，再利用去燥后的训练样本与其具有的收率标签，对初始模型进行训练，以得到指标与收率之间的关系模型。可见，本实施例中利用样本数据中的时刻属性和指标属性对样本数据进行去噪，使得用于训练关系模型的样本数据更加准确，由此使得利用这些样本数据所训练出的关系模型更加准确，从而达到提高关系模型准确率的目的。

[0102] 在一种实现方式中，去噪单元402中可以包括以下结构，如图5中所示：

[0103] 关联获得子单元421，用于获得目标指标的指标数据中的所述样本数据之间的关联关系；所述目标指标为所述至少一个指标中的任意一个；

[0104] 样本计算子单元422，用于基于所述关联关系，对所述目标指标的指标数据中的样本数据进行计算，以得到所述目标指标在所述时刻下的新的样本数据。

[0105] 另外，本实施例中的装置还可以包括以下结构单元，如图6中所示：

[0106] 模型搭建单元404，用于预先基于预设的机器学习算法搭建所述初始模型，所述初始模型为拟合模型，且所述初始模型具有至少一个与指标和收率相关的初始拟合参数。

[0107] 需要说明的是，本实施例的装置中各单元的具体实现可以参考前文中相应附图及内容，此处不再详述。

[0108] 参考图7，为本申请实施例三提供的一种电子设备的结构示意图，该电子设备可以为具有数据处理能力并能够进行模型训练的设备，如计算机或服务器等终端设备。本申请的电子设备主要用于对用于训练指标与收率之间的关系模型的样本数据进行去噪，从而达到能够提高关系模型准确率的目的。

[0109] 具体的，本实施例中的电子设备可以包括有以下结构：

[0110] 存储器701，用于存储应用程序及应用程序运行所产生的数据；

[0111] 处理器702，用于执行所述应用程序，以实现：

[0112] 获得至少一个训练样本，所述训练样本包括至少一个指标的指标数据，所述指标

数据包括其所属指标在至少一个时刻下的样本数据，所述训练样本具有预设的收率标签；基于所述样本数据对应的时刻属性和指标属性，对所述训练样本中的样本数据进行去噪；基于经过去噪的训练样本与其具有的收率标签，对初始模型进行训练，得到关系模型，所述关系模型表征所述指标与收率之间的关系。

[0113] 由以上方案可知，本申请实施例三提供的一种电子设备，在获得到各指标在各时刻下的样本数据之后，基于这些样本数据中的时刻属性和指标属性，对样本数据进行去噪，再利用去燥后的训练样本与其具有的收率标签，对初始模型进行训练，以得到指标与收率之间的关系模型。可见，本实施例中利用样本数据中的时刻属性和指标属性对样本数据进行去噪，使得用于训练关系模型的样本数据更加准确，由此使得利用这些样本数据所训练出的关系模型更加准确，从而达到提高关系模型准确率的目的。

[0114] 在一种实现方式中，处理器702在进行去噪时可以通过以下方式实现：

[0115] 获得目标指标的指标数据中的所述样本数据之间的关联关系；所述目标指标为所述至少一个指标中的任意一个；

[0116] 基于所述关联关系，对所述目标指标的指标数据中的样本数据进行计算，以得到所述目标指标在所述时刻下的新的样本数据。

[0117] 其中，所述关联关系包括所述目标指标的指标数据中的所述样本数据之间的相似度值；相应的，基于所述关联关系，对所述目标指标的指标数据中的样本数据进行计算，以得到所述目标指标在所述时刻下的新的样本数据可以通过以下方式实现：

[0118] 基于所述相似度值，确定所述目标指标的指标数据中在所述至少一个时刻中所述时刻对应的样本权重系数；基于所述样本权重系数，对所述目标指标的指标数据中的样本数据进行计算，以得到所述目标指标在所述时刻下的新的样本数据。

[0119] 在一种实现方式中，处理器702在进行去噪时也可以通过以下方式实现：

[0120] 基于所述样本数据对应的时刻属性和指标属性之间的关联关系，对所述样本数据进行更新，以使得所述训练样本中的样本数据实现去噪。

[0121] 例如，利用自注意力机制，基于所述样本数据对应的时刻属性和指标属性之间的关联关系，对所述样本数据进行更新。

[0122] 另外，所述初始模型为基于预设的机器学习算法所搭建的拟合模型，所述初始模型具有至少一个与指标和收率相关的初始拟合参数

[0123] 需要说明的是，本实施例的电子设备中处理器的具体实现可以参考前文中相应附图及内容，此处不再详述。

[0124] 以下对本实施例中的技术方案应用在制造企业中的具体示例进行举例说明，本实施例中通过对用于在收率最大化时对指标进行寻优的拟合模型中，加入基于self-attention的噪声数据处理机制，进而对模型训练中的训练样本首先进行去噪，再进行模型参数的优化训练，从而得到更加准确的拟合模型，用于指标寻优。以下对本实施例中的核心过程进行介绍：

[0125] 1)数据获取与样本构造，即获取训练样本，包括有多个指标上在多个时刻下的采样数据(样本数据)

[0126] 2)带self-attention噪声数据处理机制的拟合模型设计

[0127] 其中，由于工厂环境嘈杂等原因，用于构造样本数据的以上数据中通常存在噪声

数据，导致所构造的样本数据也含有比较多的噪声数据，对构建收率模型(即关系模型)造成不利影响。

[0128] 为此，本实施例中通过对样本数据进行去噪之后，再进行模型训练，从而提高模型的准确率。具体的，本实施例中利用self-attention结构结合样本数据在时序(时刻)和特征(指标)两个纬度上的关联信息，对传入模型的样本数据生成新的表示(新的样本数据)，该表示有效降低了原始样本数据中的噪声。

[0129] 3)模型训练

[0130] 将1)中构造的样本，带入2)中设计的带self-attention噪声处理机制的拟合模型进行训练，生成表明指标和收率之间关系的拟合关系模型。

[0131] 可见，本实施例中的技术方案中具有如下优势：

[0132] 1)使用自注意力(self-attention)机制对样本数据进行去噪，该机制可以结合时序、指标两个纬度，对数据进行平滑、扩展或修改等去噪处理，而在去噪处理时参考了更多的信息，使得噪声处理效果更好；

[0133] 2)将自注意力(self-attention)机制作为整个模型训练的一部分，嵌入到模型训练的流程中，提供了从噪声数据处理到模型训练的端到端的完整流程。避免了传统噪声数据处理方案中，噪声数据处理与模型训练不连续导致的噪声数据处理效果较差的问题。

[0134] 具体实现中，结合如图8中所示的整体框架图及图9中所示的带self-attention噪声数据处理机制的拟合模型训练逻辑示意图所示，本实施例中在使用自注意力(self-attention)机制进行噪声数据处理的工业领域高价值产品收率建模时，主要包含“数据准备”、“模型设计”、“模型训练”三个阶段，其中：

[0135] 1数据准备

[0136] 其中，该步骤主要从需要进行高价值产品收率最大化的指标寻优的目标流程制造企业的数据库(如IP21实时数据库)中获取其采集的体现产品生产装置的运行状况的指标(用:X1、X2、...、Xn等表示，如装置温度等)以及实际的高价值产品收率(用Y表示，如石化行业的汽油收率)，将这些数据构造成样本数据。如表1的样本数据所示。

[0137] 表1样本数据

[0138]

时间|指标	X1	X2	...	Xn	收率Y
T1	x1_1	X2_1	...	Xn_1	Y_1
T2	X1_2	X2_2	...	Xn_2	Y_2
T3	X1_3	X2_3	...	Xn_3	Y_3
T4	X1_4	X2_4	...	Xn_4	Y_4
...	...	...	...	...	...
TM	X1_m	X2_m	...	Xn_m	Y_m

[0139] 其中，“时间”是指标的采集的时间，高价值产品收率Y可以直接通过产品生产装置上的测量计获取(如果装置上有对应的测量点)或者通过其他测量点的值间接计算得出。

[0140] 2构建带self-attention噪声处理机制的收率拟合模型

[0141] 如图9中模型训练的逻辑层次所示：

[0142] 输入层表示各指标下各时刻对应的样本数据的输入。例如，三个指标X1、X2、X3代表连续三个时刻的样本组合，各自的三个维度(如X1:X11,X12,X13)代表每一个时刻样本的

不同指标,实际中的数量与选择的装置工艺指标数量有关,可以为其他数量的时刻;

[0143] 自注意力(self-attention)层,用于对输入层传过来的样本计算样本之间的相关性(这里对应于X1,X2,X3),基于样本之间的相关性如相似度值,重新生成每个样本对应的矢量表示(新的样本数据),如把X1的表示由原来的[X11,X12,X13]重新表示成了[Xs11,Xs12,…,Xs1m]。例如,采用前文中通过相似度值设置权重系数,并进行样本数据的技术的方式实现样本数据的新的矢量表示。该矢量表示首先综合了多个特征的信息计算了多个样本之间性关性,然后利用多个样本之间的相关性,重新生成了每个样本的矢量表示。有效利用了特征之间以及样本之间的信息,可以有效的减少噪声数据,提高拟合模型的准确性。

[0144] 隐藏层,该层是一个常规的全连接层,前接使用自注意力(self-attention)机制生成的矢量表示,后接输出层,为输出层输出中间量H11-H1n;

[0145] 输出层,该层是一个带sigmoid激活的全连接层,在H上输出收率Y。

[0146] 需要说明的是,图9中只是收率模型的一般结构,具体的实现上可以有不同的方式,如自注意力(self-attention)层、隐藏层的层数,每一层神经元的数量都可以根据实际情况进行调整,并不唯一。

[0147] 3模型训练

[0148] 在获得样本数据并构建模型的基础上,将1中准备的数据送入2中设计的模型进行训练。训练过程中对模型的模型参数进行调整优化,得到能进行准确拟和的参数组合,做为最终的拟合模型。

[0149] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置而言,由于其与实施例公开的方法相对应,所以描述的比较简单,相关之处参见方法部分说明即可。

[0150] 专业人员还可以进一步意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、计算机软件或者二者的结合来实现,为了清楚地说明硬件和软件的可互换性,在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

[0151] 结合本文中所公开的实施例描述的方法或算法的步骤可以直接用硬件、处理器执行的软件模块,或者二者的结合来实施。软件模块可以置于随机存储器(RAM)、内存、只读存储器(ROM)、电可编程ROM、电可擦除可编程ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的存储介质中。

[0152] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下,在其它实施例中实现。因此,本申请将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

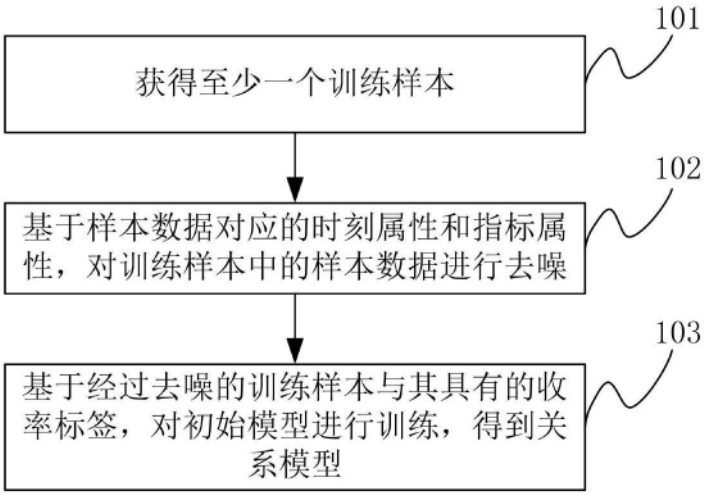

101
获得至少一个训练样本
102
基于样本数据对应的时刻属性和指标属性，对训练样本中的样本数据进行去噪
103
基于经过去噪的训练样本与其具有的收率标签，对初始模型进行训练，得到关系模型

图1

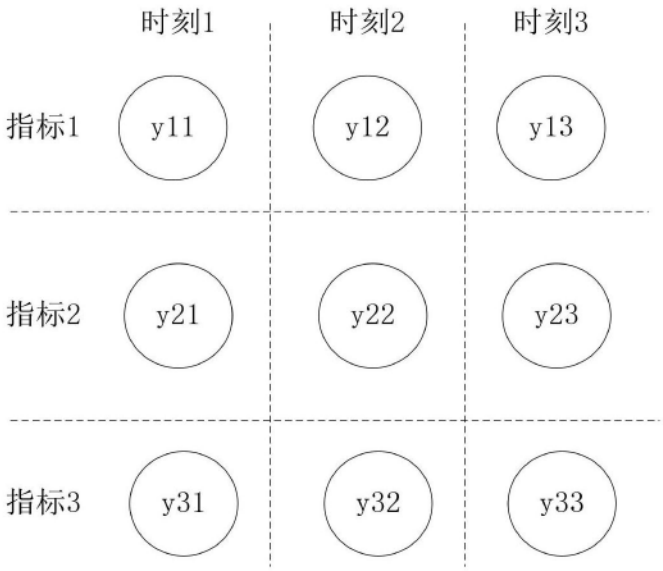

时刻1
时刻2
时刻3
指标1
y11
y12
y13
指标2
y21
y22
y23
指标3
y31
y32
y33

图2

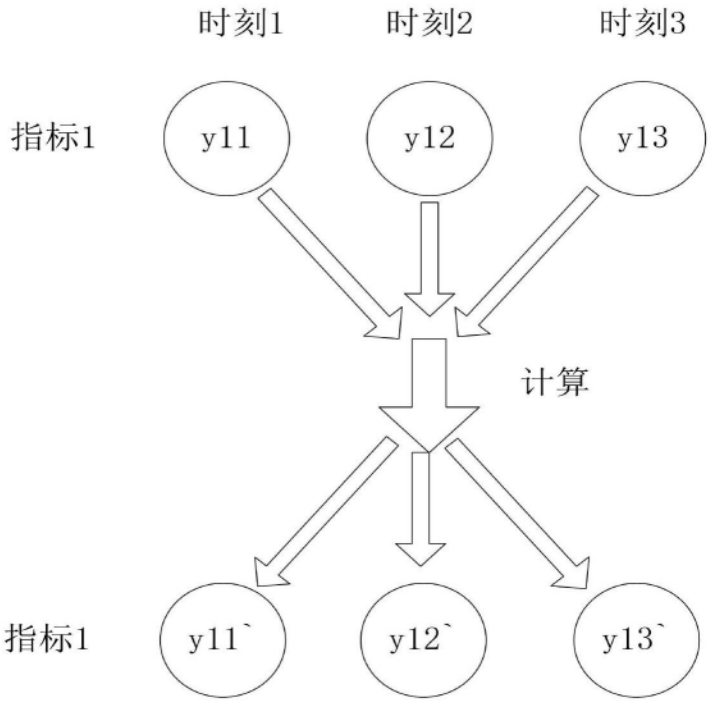

时刻1
时刻2
时刻3
指标1
y11
y12
y13
计算
指标1
y11`
y12`
y13`

图3

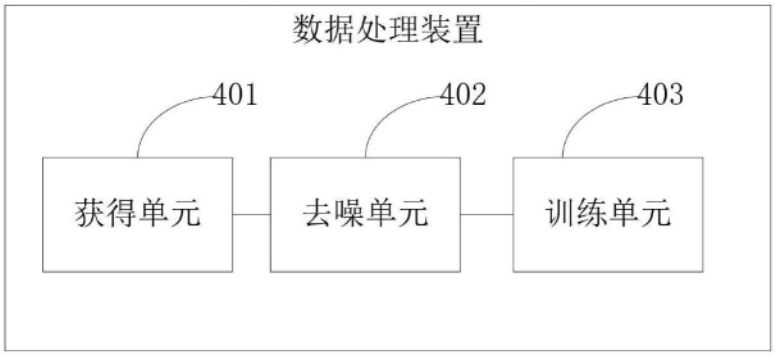

数据处理装置
401
402
403
获得单元
去噪单元
训练单元

图4

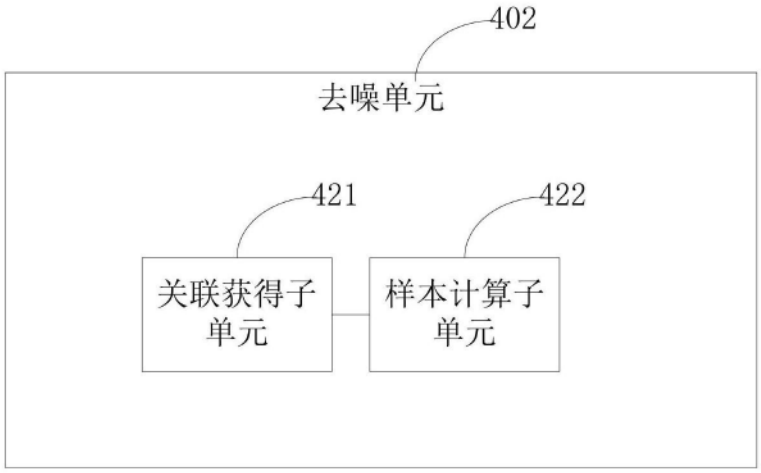

402
去噪单元
421
422
关联获得子单元
样本计算子单元

图5

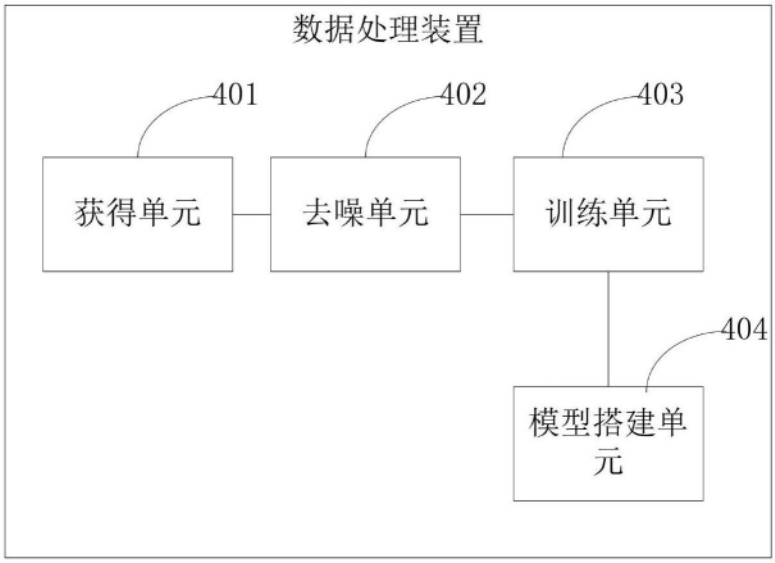

数据处理装置
401
402
403
获得单元
去噪单元
训练单元
404
模型搭建单元

图6

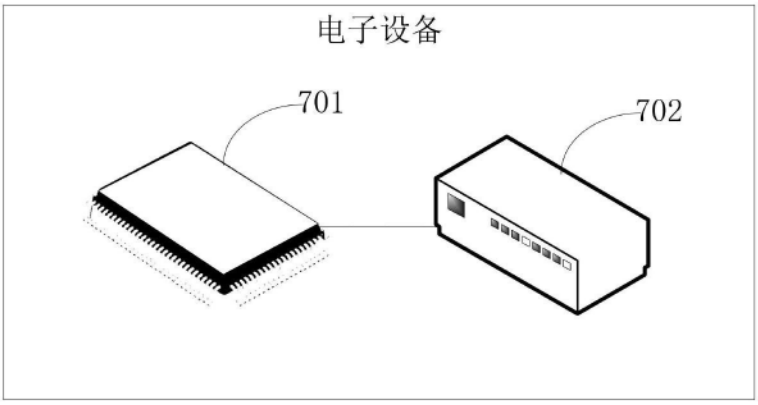

电子设备
701
702

图7

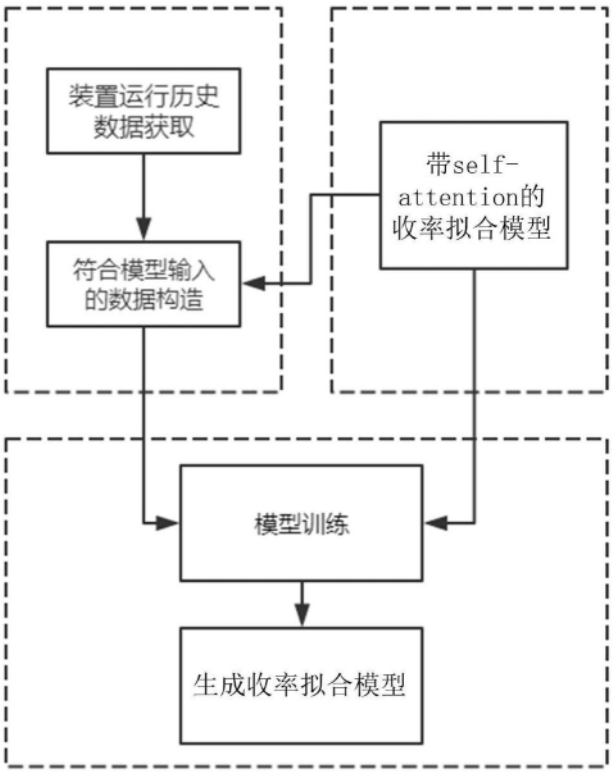

装置运行历史数据获取
带self-attention的收率拟合模型
符合模型输入的数据构造
模型训练
生成收率拟合模型

图8

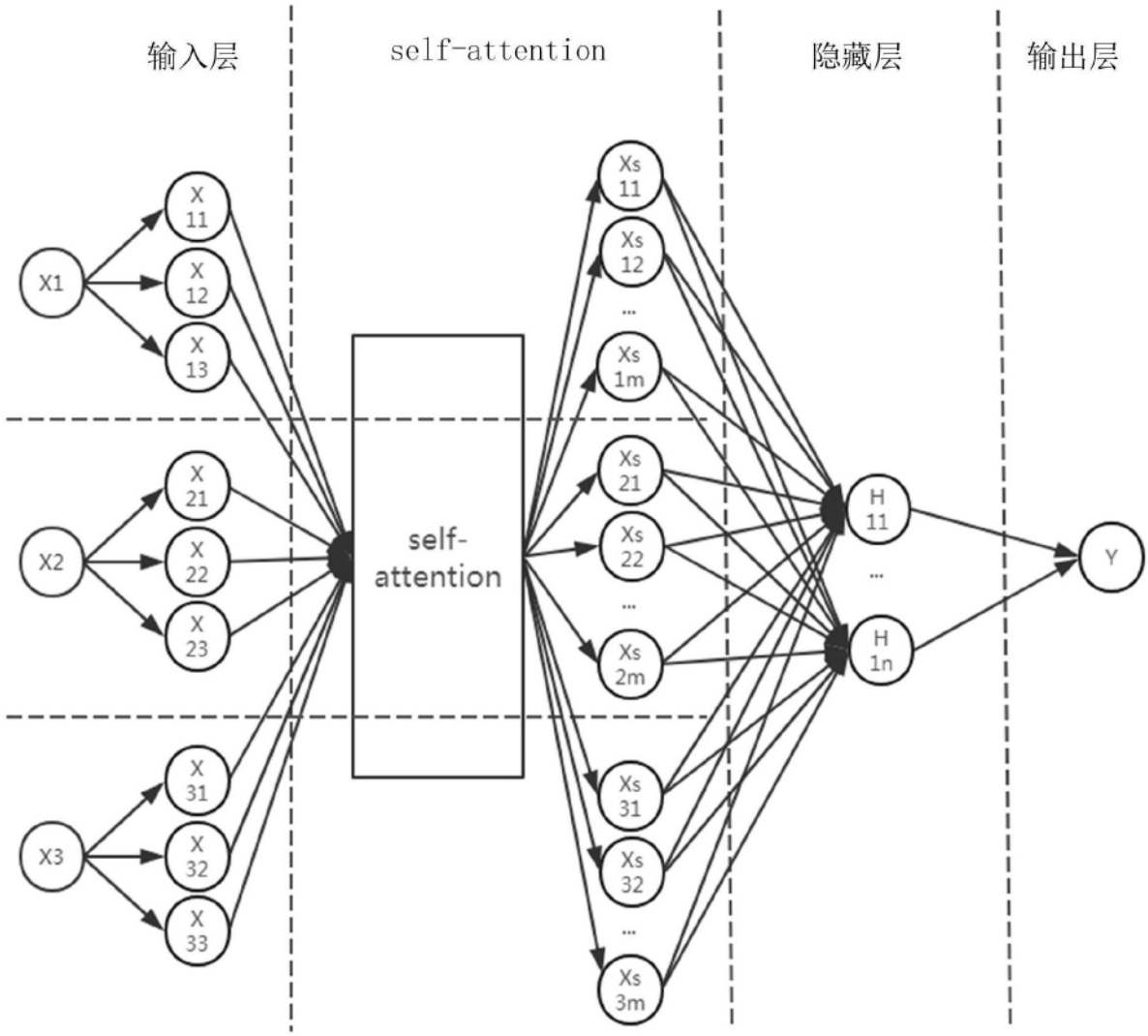

输入层
self-attention
隐藏层
输出层
X1
X2
X3
X 11
X 12
X 13
X 21
X 22
X 23
X 31
X 32
X 33
self-attention
Xs 11
Xs 12
Xs 1m
Xs 21
Xs 22
Xs 2m
Xs 31
Xs 32
Xs 3m
H 11
H 1n
Y

图9