(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112329983 B
(45)授权公告日 2024.07.26

(21)申请号 202011062341.3

(22)申请日 2020.09.30

(65)同一申请的已公布的文献号
申请公布号 CN 112329983 A

(43)申请公布日 2021.02.05

(73)专利权人 联想(北京)有限公司
地址 100085 北京市海淀区上地西路6号2幢2层201-H2-6

(72)发明人 杨帆 金继民

(74)专利代理机构 北京乐知新创知识产权代理事务所(普通合伙) 11734
专利代理师 周伟

(51)Int.Cl.
G06Q 10/0637(2023.01)
G06Q 10/0639(2023.01)
G06Q 50/02(2024.01)
G06N 3/0442(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 109993358 A,2019.07.09
CN 111160659 A,2020.05.15
杨帆 等.基于人工智能算法的催化裂化装置汽油收率预测模型的构建与分析.《石油学报(石油加工)》.2019,第35卷(第4期),807-817.

审查员 周洋

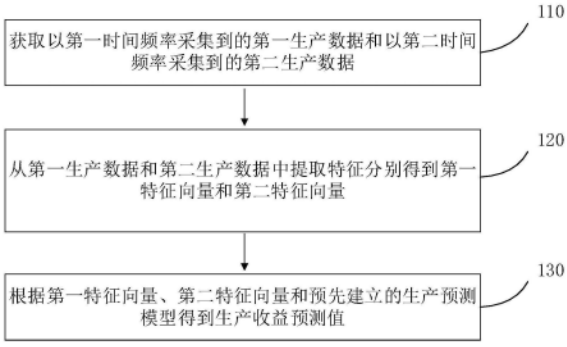
权利要求书3页 说明书11页 附图3页

(54)发明名称

一种数据处理方法及装置

(57)摘要

本发明公开了一种数据处理方法及装置,该方法包括:通过构建多处理层结构的生产预测模型,对不同时间频率采集到的生产数据进行时间对齐和信息堆叠处理,并在此基础上进行生产收益预测,其中,生产预测模型包括第一处理层、第二处理层和第三处理层,第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,第二处理层,用于对第一特征信息和第二特征信息进行合并得到第三特征信息,第三处理层,用于提取第三特征信息的时间和空间特征得到第四特征信息,并根据第四特征信息得到生产收益预测值。如此,提高了生产收益预测的准确性。



CN 112329983 B

1.一种数据处理方法,所述方法包括:

获取以第一时间频率采集到的第一生产数据和以第二时间频率采集到的第二生产数据,其中,所述第一时间频率低于所述第二时间频率,生成数据包括生产过程中的工况数据、原料配比信息、设定生产条件的配置信息;

从所述第一生产数据和所述第二生产数据中提取特征分别得到第一特征向量和第二特征向量;

根据所述第一特征向量、所述第二特征向量和预先建立的生产预测模型得到生产收益预测值,其中,所述生产预测模型包括第一处理层、第二处理层和第三处理层,

所述第一处理层,用于依据所述第一时间频率对所述第一特征向量和所述第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,所述信息叠加处理用于对所述时间对齐的时间段内的信息进行综合和累加以确保信息的完整性,

所述第二处理层,用于对所述第一特征信息和所述第二特征信息进行合并得到第三特征信息,

所述第三处理层,用于提取所述第三特征信息的时间和空间特征得到第四特征信息,并根据所述第四特征信息得到所述生产收益预测值。

2.根据权利要求1所述的方法,所述获取以第一时间频率采集到的第一生产数据和以第二时间频率采集到的第二生产数据,包括:

获取以第一时间频率采集到的第一原始数据和以第二时间频率采集到的第二原始数据;

对所述第一原始数据和所述第二原始数据进行预处理得到第一生产数据和第二生产数据,所述预处理包括缺失值、异常值和重复值的检测和处理。

3.根据权利要求1所述的方法,所述从所述第一生产数据和所述第二生产数据中提取特征分别得到第一特征向量和第二特征向量,包括:

获取与生产收益密切相关的生产特征;

根据所述与生产收益密切相关的生产特征对所述第一生产数据和所述第二生产数据进行特征数据筛选,分别得到第一特征数据和第二特征数据;

从所述第一特征数据和所述第二特征数据中提取特征得到第一特征向量和第二特征向量。

4.根据权利要求1所述的方法,所述生产预测模型还包括:

第四处理层,用于对第四特征信息进行进一步的特征提取得到第五特征信息,所述第四处理层包括标准化处理、线性变换函数和激活函数,所述激活函数包括带参数的线性整流函数。

5.根据权利要求1所述的方法,所述生产预测模型包括长短时记忆神经网络LSTM层,其中,所述长短时记忆神经网络层包括作为所述第一处理层的第一子长短时记忆神经网络层、作为所述第二处理层的连接层和作为所述第三处理层的第二子长短时记忆神经网络层。

6.一种生产预测模型的构建方法,所述方法包括:

确定生产预测模型并设置初始参数得到第一生产预测模型,所述第一生产预测模型包括第一处理层、第二处理层和第三处理层,

所述第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,所述信息叠加处理用于对所述时间对齐的时间段内的信息进行综合和累加以确保信息的完整性,

所述第二处理层,用于对所述第一特征信息和所述第二特征信息进行合并得到第三特征信息,

所述第三处理层,用于提取所述第三特征信息的时间和空间特征得到第四特征信息,并根据所述第四特征信息得到所述生产收益预测值;

获取以第一时间频率采集到的第一训练数据和以第二时间频率采集到的第二训练数据,其中,所述第一时间频率低于所述第二时间频率;

使用所述第一训练数据和所述第二训练数据对所述第一生产预测模型进行训练得到训练结果,训练数据包括生产过程中的工况数据、原料配比信息、设定生产条件的配置信息;

根据所述训练结果调整所述初始参数得到第二生产预测模型。

7.根据权利要求6所述的方法,所述使用所述第一训练数据和所述第二训练数据对所述第一生产预测模型进行训练得到训练结果,包括:

使用优化器和所述第一训练数据及所述第二训练数据对所述第一生产预测模型进行训练得到训练结果。

8.根据权利要求7所述的方法,所述优化器包括Adam优化器。

9.一种数据处理装置,所述装置包括:

生产数据获取模块,用于获取以第一时间频率采集到的第一生产数据和以第二时间频率采集到的第二生产数据,其中,所述第一时间频率低于所述第二时间频率,生成数据包括生产过程中的工况数据、原料配比信息、设定生产条件的配置信息;

特征向量获取模块,用于从所述第一生产数据和所述第二生产数据中提取特征分别得到第一特征向量和第二特征向量;

生产收益预测模块,用于根据所述第一特征向量、所述第二特征向量和预先建立的生产预测模型得到生产收益预测值,其中,所述生产预测模型包括第一处理层、第二处理层和第三处理层,

所述第一处理层,用于依据所述第一时间频率对所述第一特征向量和所述第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,所述信息叠加处理用于对所述时间对齐的时间段内的信息进行综合和累加以确保信息的完整性,

所述第二处理层,用于对所述第一特征信息和所述第二特征信息进行合并得到第三特征信息,

所述第三处理层,用于提取所述第三特征信息的时间和空间特征得到第四特征信息,并根据所述第四特征信息得到所述生产收益预测值。

10.一种生产预测模型的构建装置,所述装置包括:

初始模型构建模块,用于确定生产预测模型并设置初始参数得到第一生产预测模型,所述第一生产预测模型包括第一处理层、第二处理层和第三处理层,

所述第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,所述信息叠加处理用于对所

述时间对齐的时间段内的信息进行综合和累加以确保信息的完整性，

所述第二处理层，用于对所述第一特征信息和所述第二特征信息进行合并得到第三特征信息，

所述第三处理层，用于提取所述第三特征信息的时间和空间特征得到第四特征信息，并根据所述第四特征信息得到所述生产收益预测值；

训练数据获取模块，用于获取以第一时间频率采集到的第一训练数据和以第二时间频率采集到的第二训练数据，其中，所述第一时间频率低于所述第二时间频率，训练数据包括生产过程中的工况数据、原料配比信息、设定生产条件的配置信息；

生产预测模型训练模块，用于使用所述第一训练数据和所述第二训练数据对所述第一生产预测模型进行训练，并根据所述训练结果调整所述初始参数得到第二生产预测模型。

一种数据处理方法及装置

技术领域

[0001] 本发明涉及计算机数据处理领域,尤其涉及一种数据处理方法及装置。

背景技术

[0002] 催化裂化是一种复杂的流程制造工艺,是重油加工的最重要方法之一。目前,中国30%的柴油和70%的汽油都能过催化裂化装置加工而成,建立准确的催化裂化生产过程模型对提高原料利用率和高价值产品收率具有重要意义。

[0003] 构建催化裂化生产收率模型主要存在以下两个挑战:1)数据记录频率不一致,传感器对操作变量的记录几乎是实时的,原料油性质等指标通过人工化验记录,操作变量和原料油性质等数据的记录频率存在较大差异;2)时间延迟,由于原料油会滞留在催化裂化装置中一段时间且装置的工艺参数不断发生变化,因此,在不同的时刻,我们无法明确汽油收率受哪些工艺参数的影响。

[0004] 目前针对1)的一般处理方法主要包括降采样和平滑,但是这种方法会造成信息的损失;针对2)的一般解决方法主要为:将一段过去时间内的生产条件和当前的生产条件的串联结果作为输入变量,但是这种方法会增加模型的参数。

[0005] 因此,如何能克服现有方案的不足并解决以上问题是催化裂化生产收率预测系统中尚待解决的一个技术问题。

发明内容

[0006] 针对以上问题,本发明实施例提供了一种数据处理方法及装置。

[0007] 根据本发明实施例第一方面,一种数据处理方法,该方法包括:获取以第一时间频率采集到的第一生产数据和以第二时间频率采集到的第二生产数据,其中,第一时间频率低于第二时间频率;从第一生产数据和第二生产数据中提取特征分别得到第一特征向量和第二特征向量;根据第一特征向量、第二特征向量和预先建立的生产预测模型得到生产收益预测值,其中,生产预测模型包括第一处理层、第二处理层和第三处理层,第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,第二处理层,用于对第一特征信息和第二特征信息进行合并得到第三特征信息,第三处理层,用于提取第三特征信息的时间和空间特征得到第四特征信息,并根据第四特征信息得到生产收益预测值。

[0008] 根据本发明实施例一实施方式,获取以第一时间频率采集到的第一生产数据和以第二时间频率采集到的第二生产数据,包括:获取以第一时间频率采集到的第一原始数据和以第二时间频率采集到的第二原始数据;对第一原始数据和第二原始数据进行预处理得到第一生产数据和第二生产数据,预处理包括缺失值、异常值和重复值的检测和处理。

[0009] 根据本发明实施例一实施方式,从第一生产数据和第二生产数据中提取特征分别得到第一特征向量和第二特征向量,包括:获取与生产收益密切相关的生产特征;根据与生产收益密切相关的生产特征对第一生产数据和第二生产数据进行特征数据筛选,分别得到

第一特征数据和第二特征数据;从第一特征数据和第二特征数据中提取特征得到第一特征向量和第二特征向量。

[0010] 根据本发明实施例一实施方式,生产预测模型还包括:第四处理层,用于对第四特征信息进行进一步的特征提取得到第五特征信息,第四处理层包括标准化处理、线性变换函数和激活函数,激活函数包括带参数的线性整流函数。

[0011] 根据本发明实施例一实施方式,生产预测模型包括长短时记忆神经网络LSTM层,其中,长短时记忆神经网络层包括作为第一处理层的第一子长短时记忆神经网络层、作为第二处理层的连接层和作为第三处理层的第二子长短时记忆神经网络层。

[0012] 根据本发明实施例第二方面,一种生产预测模型的构建方法,该方法包括:确定生产预测模型并设置初始参数得到第一生产预测模型,第一生产预测模型包括第一处理层、第二处理层和第三处理层,第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,第二处理层,用于对第一特征信息和第二特征信息进行合并得到第三特征信息,第三处理层,用于提取第三特征信息的时间和空间特征得到第四特征信息,并根据第四特征信息得到生产收益预测值;获取以第一时间频率采集到的第一训练数据和以第二时间频率采集到的第二训练数据,其中,第一时间频率低于第二时间频率;使用第一训练数据和第二训练数据对第一生产预测模型进行训练得到训练结果;根据训练结果调整初始参数得到第二生产预测模型。

[0013] 根据本发明实施例一实施方式,使用第一训练数据和第二训练数据对第一生产预测模型进行训练得到训练结果,包括:选用合适的优化器;使用优化器和工业训练数据对工业产品收率预测模型进行训练得到训练结果。

[0014] 根据本发明实施例一实施方式,优化器包括Adam优化器。

[0015] 根据本发明实施例第三方面,一种数据处理装置,该装置包括:生产数据获取模块,用于获取以第一时间频率采集到的第一生产数据和以第二时间频率采集到的第二生产数据,其中,第一时间频率低于第二时间频率;特征向量获取模块,用于从第一生产数据和第二生产数据中提取特征分别得到第一特征向量和第二特征向量;生产收益预测模块,用于根据第一特征向量、第二特征向量和预先建立的生产预测模型得到生产收益预测值,其中,生产预测模型包括第一处理层、第二处理层和第三处理层,第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,第二处理层,用于对第一特征信息和第二特征信息进行合并得到第三特征信息,第三处理层,用于提取第三特征信息的时间和空间特征得到第四特征信息,并根据第四特征信息得到生产收益预测值。

[0016] 根据本发明实施例一实施方式,生产数据获取模块包括:原始数据获取子模块,用于获取以第一时间频率采集到的第一原始数据和以第二时间频率采集到的第二原始数据;预处理子模块,用于对第一原始数据和第二原始数据进行预处理得到第一生产数据和第二生产数据,预处理包括缺失值、异常值和重复值的检测和处理。

[0017] 根据本发明实施例一实施方式,特征向量获取模块包括:生产特征获取子模块,用于获取与生产收益密切相关的生产特征;特征数据筛选子模块,用于根据与生产收益密切相关的生产特征对第一生产数据和第二生产数据进行特征数据筛选,分别得到第一特征数据和第二特征数据;特征提取子模块,用于从第一特征数据和第二特征数据中提取特征得

到第一特征向量和第二特征向量。

[0018] 根据本发明实施例第三方面,一种生产预测模型的构建装置,该装置包括:初始模型构建模块,用于确定生产预测模型并设置初始参数得到第一生产预测模型,第一生产预测模型包括第一处理层、第二处理层和第三处理层,第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,第二处理层,用于对第一特征信息和第二特征信息进行合并得到第三特征信息,第三处理层,用于提取第三特征信息的时间和空间特征得到第四特征信息,并根据第四特征信息得到生产收益预测值;训练数据获取模块,用于获取以第一时间频率采集到的第一训练数据和以第二时间频率采集到的第二训练数据,其中,第一时间频率低于第二时间频率;生产预测模型训练模块,用于使用第一训练数据和第二训练数据对第一生产预测模型进行训练,并根据训练结果调整初始参数得到第二生产预测模型。

[0019] 根据本发明实施例一实施方式,生产预测模型训练模块具体用于使用优化器和工业训练数据对工业产品收率预测模型进行训练得到训练结果。

[0020] 根据本发明实施例一实施方式,优化器包括Adam优化器。

[0021] 本发明公开了一种数据处理方法及装置,该方法包括:通过构建多处理层结构的生产预测模型,对不同时间频率采集到的生产数据进行时间对齐和信息堆叠处理,并在此基础上进行生产收益预测,其中,生产预测模型包括第一处理层、第二处理层和第三处理层,第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,第二处理层,用于对第一特征信息和第二特征信息进行合并得到第三特征信息,第三处理层,用于提取第三特征信息的时间和空间特征得到第四特征信息,并根据第四特征信息得到生产收益预测值。

[0022] 由于本发明实施例在预测之前,首先对不同时间频率采集到的生产数据进行了时间对齐处理,很好地解决了数据在时间上不能准确对应的问题;其次,在进行时间对齐处理的过程中,本发明实施例对较高时间频率采集的数据进行了信息叠加而非简单采样的处理方式,因此也不存在数据损失的问题;此外,也无须将一段过去时间内的生产条件和当前的生产条件的串联结果作为输入变量。如此,不用增加模型参数,即可实现数据在时间上的同步和信息内容上的完整性,进而大大提高了生产收益预测的准确性。

[0023] 需要理解的是,本发明的教导并不需要实现上面所述的全部有益效果,而是特定的技术方案可以实现特定的技术效果,并且本发明的其他实施方式还能够实现上面未提到的有益效果。

附图说明

[0024] 通过参考附图阅读下文的详细描述,本发明示例性实施方式的上述以及其他目的、特征和优点将变得易于理解。在附图中,以示例性而非限制性的方式示出了本发明的若干实施方式,其中:

[0025] 在附图中,相同或对应的标号表示相同或对应的部分。

[0026] 图1为本发明实施例数据处理方法的实现流程示意图;

[0027] 图2为本发明实施例数据处理方法一应用预处理后的生成数据示意图;

[0028] 图3为本发明实施例数据处理方法一应用的生产预测模型结构示意图;

[0029] 图4为本发明实施例生产预测模型的构建方法的实现流程示意图；

[0030] 图5为本发明实施例本发明实施例数据处理装置的组成结构示意图；

[0031] 图6为本发明实施例本发明实施例生产预测模型的构建装置的组成结构示意图。

具体实施方式

[0032] 为使本发明的目的、特征、优点能够更加的明显和易懂，下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而非全部实施例。基于本发明中的实施例，本领域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，都属于本发明保护的范围。

[0033] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0034] 此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或隐含地包括至少一个该特征。在本发明的描述中，“多个”的含义是两个或两个以上，除非另有明确具体的限定。

[0035] 图1示出了本发明实施例数据处理方法的实现流程。参考图1，该方法包括：操作110，获取以第一时间频率采集到的第一生产数据和以第二时间频率采集到的第二生产数据，其中，第一时间频率低于第二时间频率；操作120，从第一生产数据和第二生产数据中提取特征分别得到第一特征向量和第二特征向量；操作130，根据第一特征向量、第二特征向量和预先建立的生产预测模型得到生产收益预测值，其中，生产预测模型包括第一处理层、第二处理层和第三处理层，第一处理层，用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理，分别得到第一特征信息和第二特征信息，第二处理层，用于对第一特征信息和第二特征信息进行合并得到第三特征信息，第三处理层，用于提取第三特征信息的时间和空间特征得到第四特征信息，并根据第四特征信息得到生产收益预测值。

[0036] 在操作110中，生产数据主要指生产过程中的各种工况数据、原料配比信息、设定生产条件的配置信息等影响生产收益的各种数据。其中，有些数据是通过实时采集设备，例如，摄像头、传感器、温度或湿度仪表等，以较高的时间频率实时采集的来获取的，比如生产过程中的各种工况数据；而有些数据则是通过人工化验记录的来获取，例如，原料油性质等指标等，这些数据的采集频率就比实时采集工况数据的采集频率低很多。

[0037] 在操作120中，提取特征主要指从生产数据中提取对生产收益预测其决定作用的信息，并将这些信息转换为计算机可识别和运算的特征向量。这些特征向量通常为描述某些特征的n阶矩阵。

[0038] 在提取特征得到特征向量时，通常是将一次采集得到的生产数据转换为一个特征向量，因而，在相同的时段内，数据采集频率高的生产数据转换得到的特征向量数量要比数

据采集频率低的生产数据转换得到的特征向量数量多。例如,某一种生产数据A的采集频率是每10分钟采集一次,而另一种生产数据B的采集频率是每20分钟采集一次。则在1个小时内,生产数据A在提取特征后会得到6个特征向量,而生成数据B在提取特征后会得到3个特征向量。

[0039] 在本实施方式中,并不限定特征提取所采用的具体算法或实现方式,实施者可采用任意适用的算法或实现方式。

[0040] 在操作130中,生产收益预测值主要指用于衡量生产收益的相关指标值,例如,生产收益率、成本利润率、净产值率、劳动生产率等。

[0041] 生产预测模型为多层结构的模型,包括多个处理层,其中每一处理层都对不同频率采集的数据进行特定的处理以实现特征数据在时间线上的对齐和全部信息的叠加。

[0042] 第一处理层所进行的时间对齐处理,主要指按时间段划分信息,将同一时间段内的信息划分到一组进行统一处理,实现生产数据在时间线上保持一致和同步。而时间段的最小取值通常是不同采集频率每两次采集间隔的最小公倍数。例如,某一种生产数据A的采集频率是每10分钟采集一次,而另一种生产数据B的采集频率是每20分钟采集一次,则可以使用20分钟、40分钟等作为进行对齐时间处理所采用的时间段。

[0043] 第一处理层所进行的信息叠加处理主要是指对时间对齐处理所采用的时间段内的信息进行综合和累加以确保信息的完整性,而不是简单地选取某一个时间点的信息作为代表。

[0044] 需要说明的是,第一处理层是对第一特征向量和第二特征向量分别进行处理的,得到的分别代表第一生产数据特征的特征信息和代表第人生产数据特征的特征信息。特征信息泛指可以描述某些特征的相关信息,例如,对特征向量进行特定运算或转换而得到的,新的特征向量或某一数值。

[0045] 这一处理层为后续信息的进一步合并和特征的进一步提取提供了数据基础,并奠定了时间同步性和数据完整性的基础。

[0046] 第二处理层则主要用于将第一特征信息和第二特征信息进行合并得到第三特征信息。通常对于计算机模型,特别是神经网络模型来说,同时对多种类型的特征信息进行运算的复杂度比对单一类型的特征信息进行运算的复杂度要高得多。因此,大多数计算机模型都会将各种信息通过特定算法进行合并得到一种综合了各个特征的综合特征信息

[0047] 在第三处理层所进行的处理,是通过进一步提取第三特征信息的时间和空间特征以发现数据之间在时间上和空间上的联系。这一处理,可以尽可能根据需要保留之前一段时间内的历史数据,可以更好地解决生产过程中可能发生的工况对产收率产生影响会有所延迟的问题。

[0048] 根据本发明实施例一实施方式,获取以第一时间频率采集到的第一生产数据和以第二时间频率采集到的第二生产数据,包括:获取以第一时间频率采集到的第一原始数据和以第二时间频率采集到的第二原始数据;对第一原始数据和第二原始数据进行预处理得到第一生产数据和第二生产数据,预处理包括缺失值、异常值和重复值的检测和处理。

[0049] 其中,原始数据主要指实时采集设备或人工输入的数据。由于采集设备故障,或人为录入错误会产生一些缺陷数据,比如出现缺失值、异常值等,而这些数据缺陷会破坏数据的一致性和完整性,最终导致预测偏差,因此,需要进行一些预处理以保证数据的一致性和

完整性。例如,删除或补齐缺失数据,删除或替换异常值等等。

[0050] 此外,在采集到的众多数据中,也可能存在通过不同传感设备获取的相同意义的数据,例如,生产设备数据中的炉温和生产条件数据中的生产温度就可能是同一温度且意义相同。此时,保留两份重复数据并不会提升预测结果的准确性,反而增加了预测过程中的计算量和复杂度,因此,移除这些重复数据可以简化计算过程、缩短预算时间。

[0051] 根据本发明实施例一实施方式,从第一生产数据和第二生产数据中提取特征分别得到第一特征向量和第二特征向量,包括:获取与生产收益密切相关的生产特征;根据与生产收益密切相关的生产特征对第一生产数据和第二生产数据进行特征数据筛选,分别得到第一特征数据和第二特征数据;从第一特征数据和第二特征数据中提取特征得到第一特征向量和第二特征向量。

[0052] 对于生产工艺较为复杂的生产过程,采集到的生产数据包括各方面的计量指标值,这些指标对应的生产特征甚至会达到上千量级,此时,如果不加以筛选,则预测生产收益所要进行的计算量是非常巨大的,甚至难以实现。

[0053] 因此,需要进行生产特征与生产收益之间的相关性分析并根据相关性的强弱来进行特征筛选。例如,可以通过皮尔逊相关系数来度量生产特征与生产收益之间的相关性,并选择相关性较高的生产特征。皮尔逊相关系数的计算公式如下所示:

[0054] $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

[0055] 其中,n为样本数,x_i和y_i为两个变量的观测值,$\bar{x}$和$\bar{y}$为两个变量的均值,r为计算出的两个变量相关系数。

[0056] 根据本发明实施例一实施方式,生产预测模型还包括:第四处理层,用于对第四特征信息进行进一步的特征提取得到第五特征信息,第四处理层包括标准化处理、线性变换函数和激活函数,激活函数包括带参数的线性整流函数。

[0057] 在本实施方式中,第四处理层对应于神经网络中的全连接层,用于将之前提取到的特征信息进行加权求和以得到最后的预测结果。

[0058] 根据本发明实施例一实施方式,生产预测模型包括长短时记忆神经网络(LSTM)层,其中,长短时记忆神经网络层包括作为第一处理层的第一子长短时记忆神经网络层、作为第二处理层的连接层和作为第三处理层的第二子长短时记忆神经网络层。

[0059] 在本实施方式中,生产预测模型是基于长短时记忆神经网络模型构建的,且其中的长短时记忆神经网络层又分为两个子长短时记忆神经网络层和一个连接层。第一个子长短时记忆神经网络层用于时间对齐和信息叠加,连接层用来合并以不同时间频率采集到的信息,而第二个子长短时记忆神经网络层则用于提取合并后信息的时间和空间特征。

[0060] 长短时记忆神经网络可以记忆不定时间长度的数值,且可以通过忘记门决定那些信息需要被记忆,哪些信息需要被忘记,特别适合于处理和预测时间序列中间隔和延迟非常长的重要事件。利用这一特性,可以通过建立两个时间长度不同的长短时记忆神经网络,分别对不同时间频率采集的信息进行时间对齐和信息叠加处理,再通过一个连接层进行拼接就可以实现时间上的同步性和信息的完整性。由于长短时记忆神经网络可以保留之前的历史信息,也可以更好地解决生产过程中可能发生的工况对产收率产生影响会有所延迟的

问题。

[0061] 下面就以本发明实施例的一个具体应用,进行示例性说明。

[0062] 在这一应用中,生产数据的采集主要是从石化行业的集散控制系统和实验室信息管理系统中获取能够反映生产条件的指标,例如,催化裂化装置的温度和压强指标,原料油和催化剂的性质等,而其要预测的产品收益,主要是石化行业的汽油收率。

[0063] 其中,从石化行业的集散控制系统获取的主要是实时采集到的工况数据,采集频率较高;而从实验室信息管理系统获取的主要是原料配比信息等,主要靠人工实验得出并通过人工录入系统的,其采集频率较低,并具有一定的时间延迟。

[0064] 表1就给出了这一应用的一个示例性数据:

[0065]

Xl 1	Xl 2	...	Xl m	Xd1	...	Xdn	Y
xl 1_1	Xl 2_1	...	Xl m_1	Xd1_ 1_1	...	Xdn_ 1_1	Y_ 1
				Xd1_ 1_2	...	Xdn_ 1_2	
				...	...	...	
				Xd1_ 1_j	...	Xdn_ 1_j	
Xl 1_2	Xl 1_2	...	Xl m_2	Xd1_ 2_1	...	Xdn_ 2_1	Y_2
				Xd1_ 2_2	...	Xdn_ 2_2	
				...	...	...	
				Xd1_ 2_j	...	Xdn_ 2_j	
...	...	...	...	...	...	...	...
Xl 1_N	Xl 2_N	...	Xl m_N	Xd1_ N_1	...	Xdn_ N_1	Y_ N
				Xd1_ N_2	...	Xdn_ N_2	
				...	...	...	
				Xd1_ N_j	...	Xdn_ N_j	

[0066] 表1

[0067] 其中,m表示低记录频率的指标数量,n表示较高记录频率的指标数量,Xlm_N表示第m个指标(较低记录频率的指标)的第N个记录值,Xdn N_j表示第n个指标(较高记录频率的指标)在第N个时间周期内的第j个记录值,Y_N表示产品收率在第N个时间周期内的记录

值。

[0068] 之后,对上述数据中的缺失数据、异常数据和重复数据进行预处理,得到如图2所示的数据,其中,用虚线框出的数据是同一时段,需要进行时间对齐处理的生产数据。

[0069] 然后,通过相关性分析进行特征筛选和特征提取得到第一特征向量X^l和第二特征向量X^d,其中,X^l表示低采样频率的生产数据,X^d表示高采样频率的生产输入。

[0070] 之后,将上述特征向量输入到如图3所示的生产预测模型中就可以获得石油收率的预测了。如图3所示,该生产预测模型是基于多层的神经网络模型,包括:输入层301、长短时记忆神经网络层302、全连接层303和输出层404,$\hat{y}$表示模型的输出,其中,长短时记忆神经网络层302又分为第一长短时记忆神经网络层3021、连接层3022和第二长短时记忆神经网络层3023。

[0071] 输入层301,主要用于不同采样频率变量的输入,如图3中所示,X^l表示低采样频率的输入输入变量,X^d表示高采样频率的输入变量。

[0072] 长短时记忆神经网络层302,主要用于不同采样频率数据的时间对齐和特征提取。第一长短时记忆神经网络层3021由两个长短时记忆神经网络层组成,这两个长短时记忆神经网络层分别用于处理不同采样频率的样本数据,例如,在N个时间周期内,第一长短时记忆神经网络层3021对x_N^l和$(x_{12N-11}^d, x_{12N-10}^d, \cdots, x_{12N}^d)$处理后得到$h_N^l$和$h_{12N}^d$然,其中,$h_{12N}^d$叠加了$x_{12N-11}^d, x_{12N-10}^d, \cdots, x_{12N}^d$的信息;然后,连接层3022通过对$h_N^l$和$h_{12N}^d$进行拼接,得到$Concat(h_i^l, h_{12i}^d)$,拼接结果可表示为在N个时间周期内输入变量的特征信息;最后,第二长短时记忆神经网络层3023对拼接结果进行处理,提取时间和空间特征。

[0073] 全连接层303,由1个标准层和2个线性变换及线性整流函数(prelu)组成,主要用于特征的转换、加权综合和分类。

[0074] 输出层304,主要输出产品的收率,例如,在N个时间周期内,产品收率的预测值为$\hat{y}_N$。

[0075] 需要说明的是,上述应用仅为本发明实施例可应用的一个场景之一,是示例性说明,而非对本发明实施例应用场景的限定。本发明实施例还可应用于类似的、其他行业或产品的、其他生产收益预测。

[0076] 根据本发明实施例第二方面,一种生产预测模型的构建方法,如图4所示,该方法包括:操作410,确定生产预测模型并设置初始参数得到第一生产预测模型,第一生产预测模型包括第一处理层、第二处理层和第三处理层,第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,第二处理层,用于对第一特征信息和第二特征信息进行合并得到第三特征信息,第三处理层,用于提取第三特征信息的时间和空间特征得到第四特征信息,并根据第四特征信息得到生产收益预测值;获取以第一时间频率采集到的第一训练数据和以第二时间频率采集到的第二训练数据,其中,第一时间频率低于第二时间频率;操作420,使用第一训练数据和第二训练数据对第一生产预测模型进行训练得到训练结果;操作430,根据训练结果调整初始参数得到第二生产预测模型。

[0077] 在操作410中,创建生产预测模型主要包括选取合适的算法并设置初始参数,这些

主要取决于具体应用场景的产品生产特性,本发明实施例并不对其具体算法或初始参数的设置进行限定。关于模型结构与在操作130所描述的生产预测模型的结构相同,各层的用途和所代表的含义也相同,相关细节,请参照前文描述,在此不再赘述。

[0078] 在操作420中,第一训练数据和第二训练数据,也与实际预测中所使用的生产数据类似,只是多了标注信息以设置期待的生产收益值。在实际应用中,可以收益生产数据的历史数据和这些数据对应的实际生产收益值来构建训练数据。此处的模型训练过程与常用的模型训练过程无异,故不再赘述。

[0079] 根据本发明实施例一实施方式,使用第一训练数据和第二训练数据对第一生产预测模型进行训练得到训练结果,包括:选用合适的优化器;使用优化器和工业训练数据对工业产品收率预测模型进行训练得到训练结果。

[0080] 根据本发明实施例一实施方式,优化器包括Adam优化器。

[0081] 根据本发明实施例第三方面,一种数据处理装置,如图5所示,该装置50包括:生产数据获取模块501,用于获取以第一时间频率采集到的第一生产数据和以第二时间频率采集到的第二生产数据,其中,第一时间频率低于第二时间频率;特征向量获取模块502,用于从第一生产数据和第二生产数据中提取特征分别得到第一特征向量和第二特征向量;生产收益预测模块503,用于根据第一特征向量、第二特征向量和预先建立的生产预测模型得到生产收益预测值,其中,生产预测模型包括第一处理层、第二处理层和第三处理层,第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,第二处理层,用于对第一特征信息和第二特征信息进行合并得到第三特征信息,第三处理层,用于提取第三特征信息的时间和空间特征得到第四特征信息,并根据第四特征信息得到生产收益预测值。

[0082] 根据本发明实施例一实施方式,生产数据获取模块501包括:原始数据获取子模块,用于获取以第一时间频率采集到的第一原始数据和以第二时间频率采集到的第二原始数据;预处理子模块,用于对第一原始数据和第二原始数据进行预处理得到第一生产数据和第二生产数据,预处理包括缺失值、异常值和重复值的检测和处理。

[0083] 根据本发明实施例一实施方式,特征向量获取模块502包括:生产特征获取子模块,用于获取与生产收益密切相关的生产特征;特征数据筛选子模块,用于根据与生产收益密切相关的生产特征对第一生产数据和第二生产数据进行特征数据筛选,分别得到第一特征数据和第二特征数据;特征提取子模块,用于从第一特征数据和第二特征数据中提取特征得到第一特征向量和第二特征向量。

[0084] 根据本发明实施例第三方面,一种生产预测模型的构建装置,如图6所示,该装置60包括:初始模型构建模块601,用于确定生产预测模型并设置初始参数得到第一生产预测模型,第一生产预测模型包括第一处理层、第二处理层和第三处理层,第一处理层,用于依据第一时间频率对第一特征向量和第二特征向量进行时间对齐和信息叠加处理,分别得到第一特征信息和第二特征信息,第二处理层,用于对第一特征信息和第二特征信息进行合并得到第三特征信息,第三处理层,用于提取第三特征信息的时间和空间特征得到第四特征信息,并根据第四特征信息得到生产收益预测值;训练数据获取模块602,用于获取以第一时间频率采集到的第一训练数据和以第二时间频率采集到的第二训练数据,其中,第一时间频率低于第二时间频率;生产预测模型训练模块603,用于使用第一训练数据和第二训

练数据对第一生产预测模型进行训练,并根据训练结果调整初始参数得到第二生产预测模型。

[0085] 根据本发明实施例一实施方式,生产预测模型训练模块603具体用于使用优化器和工业训练数据对工业产品收率预测模型进行训练得到训练结果。

[0086] 根据本发明实施例一实施方式,优化器包括Adam优化器。

[0087] 这里需要指出的是:以上针对数据处理的装置实施例的描述和以上针对生产预测模型的构建装置的描述,与前述方法实施例的描述是类似的,具有同前述方法实施例相似的有益效果,因此不做赘述。对于本发明对数据处理装置实施例的描述和对生产预测模型的构建装置实施例的描述尚未披露的技术细节,请参照本发明前述方法实施例的描述而理解,为节约篇幅,因此不再赘述。

[0088] 需要说明的是,在本文中,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、物品或者装置不仅包括那些要素,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、物品或者装置所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括该要素的过程、方法、物品或者装置中还存在另外的相同要素。

[0089] 在本申请所提供的几个实施例中,应该理解到,所揭露的设备和方法,可以通过其它的方式实现。以上所描述的设备实施例仅仅是示意性的,例如,单元的划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式,如:多个单元或组件可以结合,或可以集成到另一个装置,或一些特征可以忽略,或不执行。另外,所显示或讨论的各组成部分相互之间的耦合、或直接耦合、或通信连接可以是通过一些接口,设备或单元的间接耦合或通信连接,可以是电性的、机械的或其它形式的。

[0090] 上述作为分离部件说明的单元可以是、或也可以不是物理上分开的,作为单元显示的部件可以是、或也可以不是物理单元;既可以位于一个地方,也可以分布到多个网络单元上;可以根据实际的需要选择其中的部分或全部单元来实现本实施例方案的目的。

[0091] 另外,在本发明各实施例中的各功能单元可以全部集成在一个处理单元中,也可以是各单元分别单独作为一个单元,也可以两个或两个以上单元集成在一个单元中;上述集成的单元既可以利用硬件的形式实现,也可以利用硬件加软件功能单元的形式实现。

[0092] 本领域普通技术人员可以理解:实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成,前述的程序可以存储于计算机可读取存储介质中,该程序在执行时,执行包括上述方法实施例的步骤;而前述的存储介质包括:移动存储介质、只读存储器(Read Only Memory,ROM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0093] 或者,本发明上述集成的单元如果以软件功能模块的形式实现并作为独立的产品销售或使用时,也可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明实施例的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机、服务器、或者网络设备等)执行本发明各个实施例方法的全部或部分。而前述的存储介质包括:移动存储介质、ROM、磁碟或者光盘等各种可以存储程序代码的介质。

[0094] 以上,仅为本发明的具体实施方式,但本发明的保护范围并不局限于此,任何熟悉本技术领域的技术人员在本发明揭露的技术范围内,可轻易想到变化或替换,都应涵盖在

本发明的保护范围之内。因此,本发明的保护范围应以权利要求的保护范围为准。

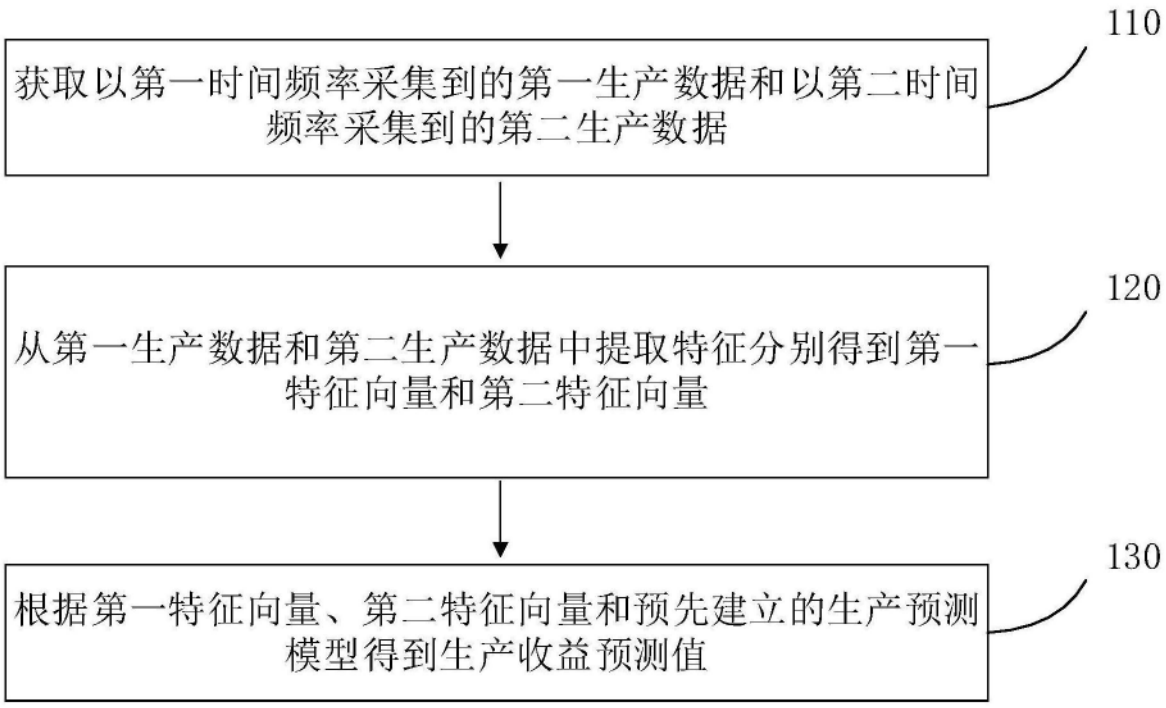


图1

记录频率低的数据

$$\underset{(N*m)}{X^{l}}=\begin{pmatrix} x_{1,1}^{l} & \cdots & x_{i-1,1}^{l} & x_{i,1}^{l} & \cdots & x_{\mathrm{N},1}^{l} \\ x_{1,2}^{l} & \cdots & x_{i-1,2}^{l} & x_{i,2}^{l} & \cdots & x_{\mathrm{N},2}^{l} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{1,m}^{l} & \cdots & x_{i-1,m}^{l} & x_{i,m}^{l} & \cdots & x_{\mathrm{N},\mathrm{m}}^{l} \end{pmatrix}$$

记录频率高的数据

$$\underset{(12N*n)}{X^{\mathrm{d}}}=\begin{pmatrix} x_{1,1}^{d} & x_{2,1}^{d} & \cdots & x_{12,1}^{d} & \cdots & x_{12i-12,1}^{d} & x_{12i-11,1}^{d} & \cdots & x_{12i,1}^{d} & x_{12i+1,1}^{d} & \cdots & x_{12N,1}^{d} \\ x_{1,2}^{d} & x_{2,2}^{d} & \cdots & x_{2,n}^{d} & \cdots & x_{12i-12,2}^{d} & x_{12i-11,2}^{d} & \cdots & x_{12i,2}^{d} & x_{12i+1,2}^{d} & \cdots & x_{12N,2}^{d} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{1,n}^{d} & x_{12,2}^{d} & \cdots & x_{12,n}^{d} & \cdots & x_{12i-12,n}^{d} & x_{12i-11,n}^{d} & \cdots & x_{12i,n}^{d} & x_{12i+1,n}^{d} & \cdots & x_{12N,n}^{d} \end{pmatrix}$$

图2

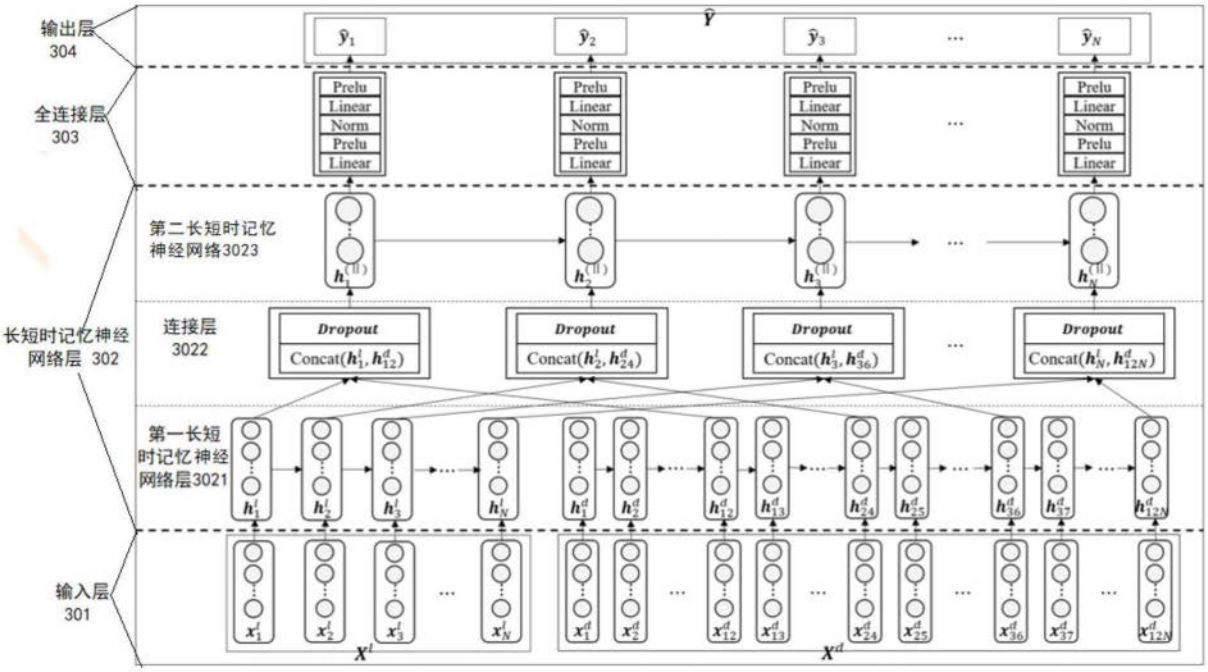

输出层
304
全连接层
303
长短时记忆神经网络层 302
连接层
3022
输入层
301
Prelu
Linear
Norm
Prelu
Linear
Dropout

图3

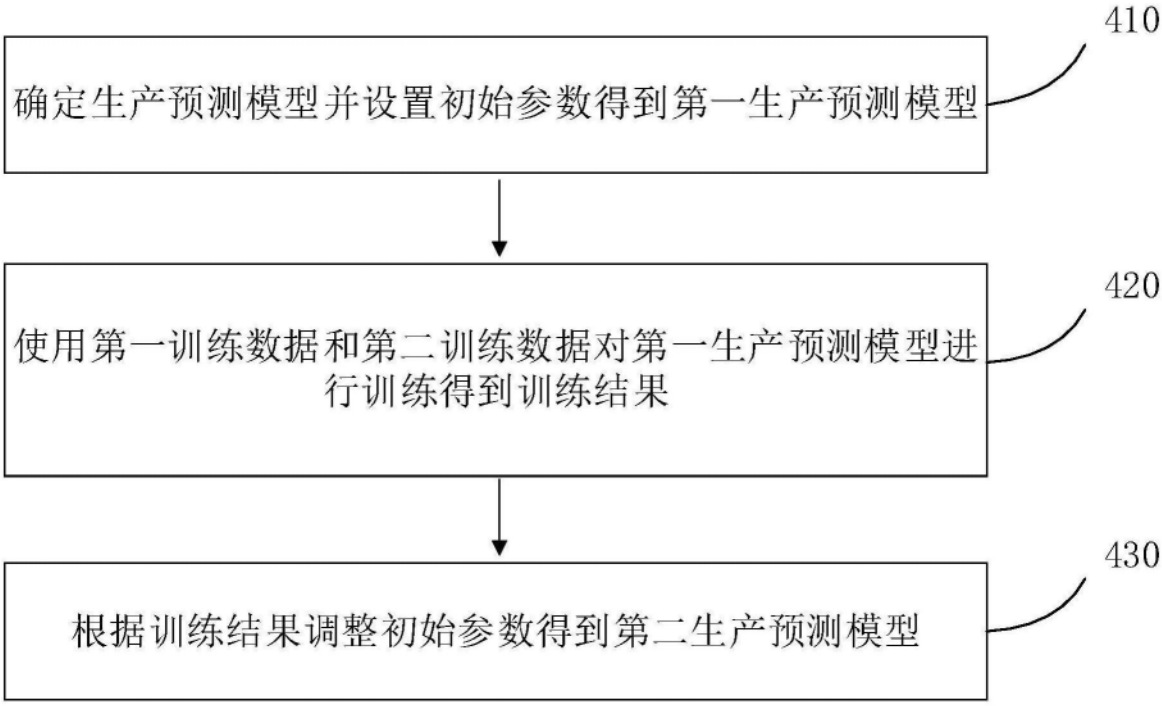

410
确定生产预测模型并设置初始参数得到第一生产预测模型
420
使用第一训练数据和第二训练数据对第一生产预测模型进行训练得到训练结果
430
根据训练结果调整初始参数得到第二生产预测模型

图4

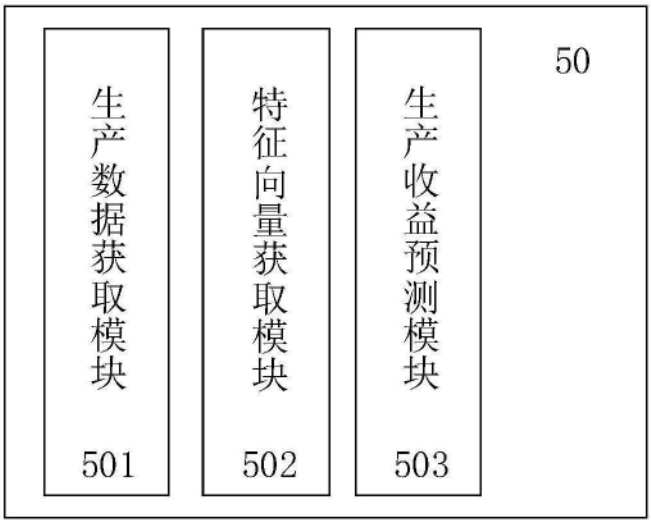

50
生产数据获取模块
501
特征向量获取模块
502
生产收益预测模块
503

图5

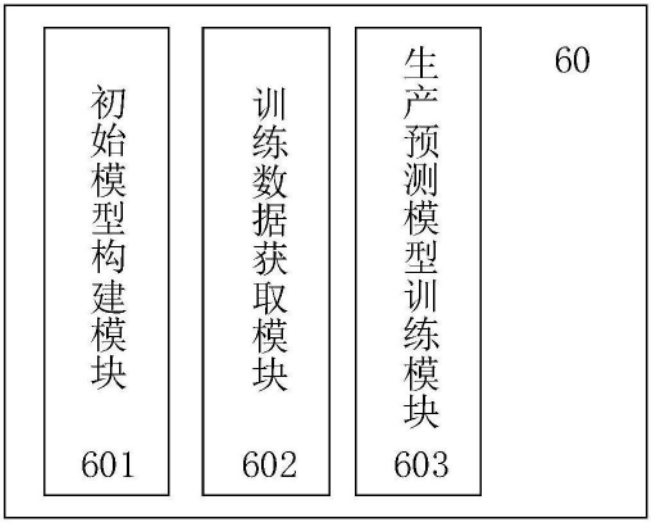

60
初始模型构建模块
601
训练数据获取模块
602
生产预测模型训练模块
603

图6