(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111833012 B
(45)授权公告日 2024.06.21

(21)申请号 202010566262.X

(22)申请日 2020.06.19

(65)同一申请的已公布的文献号
申请公布号 CN 111833012 A

(43)申请公布日 2020.10.27

(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 戴超男 杨帆 金继民 汪洁

(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 10/10(2023.01)
G06F 18/23213(2023.01)

(56)对比文件
CN 105784340 A,2016.07.20
CN 109409407 A,2019.03.01
CN 110674892 A,2020.01.10

审查员 舒霏霏

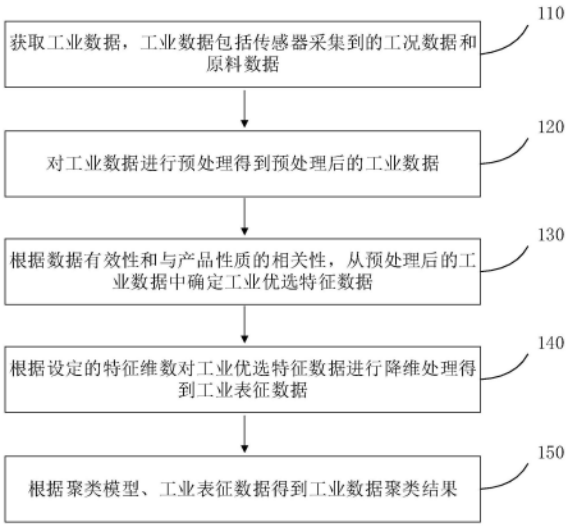
权利要求书2页 说明书9页 附图2页

(54)发明名称

一种工业数据处理方法和装置

(57)摘要

本发明实施例公开了一种工业数据处理方法和装置。该方法包括:首先,通过传感器实时采集工况数据并结合原料数据形成工业数据;之后,对工业数据进行预处理;然后,根据数据有效性和与产品性质的相关性,从预处理后的工业数据中确定工业优选特征数据;再根据设定的特征维数对工业优选特征数据进行降维处理得到工业表征数据;随后,利用聚类模型、工业表征数据得到工业数据聚类结果。如此,紧密结合产品性质对工业优选特征数据进行层层筛选,得到更具代表性的工业表征数据,并根据这些表征数据和聚类模型得到的低维聚类结果,使工业数据代表的工况更具代表性和可视性,从而为收率寻优提供了重要参考依据。



CN 111833012 B

1.一种工业数据处理方法,所述方法包括:

获取工业数据,所述工业数据包括传感器采集到的工况数据和原料数据;

对所述工业数据进行预处理得到预处理后的工业数据;

根据数据有效性和与产品性质的相关性,从所述预处理后的工业数据中确定工业优选特征数据;

根据设定的特征维数对所述工业优选特征数据进行降维处理得到工业表征数据;

根据聚类模型、所述工业表征数据得到工业数据聚类结果;

所述根据数据有效性和与产品性质的相关性,从所述预处理后的工业数据中确定工业优选特征数据,包括:获取所述预处理后的工业数据的所有特征得到特征全集;对所述特征全集进行第一轮特征筛选得到第一特征集合,所述第一轮特征筛选包括判断所述特征全集中每一特征是否唯一且有效,若是,则保留相应特征,若否,则删除或合并相应特征;对所述第一特征全集进行第二轮特征筛选得到第二特征集合,所述第二轮特征筛选包括判断所述第一特征集合中每一特征是否与产品性质密切相关,若是,则保留相应特征,若否,则删除相应特征;从所述预处理后的工业数据中获取与所述第二特征集合中的特征对应的数据作为所述工业优选特征数据;

所述判断所述第一特征集合中每一特征是否与产品性质密切相关,包括:使用皮尔逊相关系数计算所述第一特征集合中每一特征与产品性质的相关性;判断所述相关性是否大于相关性阈值,若大于,则所述特征与产品性质密切相关。

2.根据权利要求1所述的方法,在所述根据设定的特征维数对所述工业优选特征数据进行降维处理得到工业表征数据之前,所述方法还包括:

根据经验值设定特征维数,或

根据降维效果设定特征维数。

3.根据权利要求2所述的方法,所述根据降维效果设定特征维数,包括:

确定至少两个待定特征维数;

计算使用每一待定特征维数进行特征降维后得到的数据方差;

从所述至少两个待定特征维数中选取数据方差最高的待定特征维数设定特征维数。

4.根据权利要求1所述的方法,在所述根据聚类模型、所述工业表征数据得到工业数据聚类结果之前,所述方法还包括:

根据聚类效果确定聚类算法和聚类模型;

使用历史工业数据作为样本数据,对所述聚类模型进行训练和调优。

5.根据权利要求4所述的方法,所述根据聚类效果确定聚类算法和聚类模型,包括:

选取至少两个待定聚类算法;

根据聚类效果从所述至少两个待定聚类算法中确定一个聚类效果较好的待定聚类算法作为聚类算法;

根据所述聚类算法建立聚类模型。

6.根据权利要求5所述的方法,所述根据聚类效果确定聚类算法和聚类模型,包括:

使用k-means算法作为聚类算法;

根据聚类效果确定所述k-means算法中要使用的簇集数量;

根据所述k-means算法和所述簇集数量建立聚类模型。

7.根据权利要求1所述的方法,所述设定的特征维数小于等于3,相应地,在所述根据聚类模型、所述工业表征数据得到工业数据聚类结果之后,所述方法还包括:

以图形方式展示所述工业数据聚类结果以便为收率调优决策提供支持。

8.一种工业数据处理装置,所述装置包括:

数据获取模块,用于获取工业数据,所述工业数据包括传感器采集到的工况数据和原料数据;

数据预处理模块,用于对所述工业数据进行预处理得到预处理后的工业数据;

优选特征确定模块,用于根据数据有效性和与产品性质的相关性,从所述预处理后的工业数据中确定工业优选特征数据;

特征数据降维模块,用于根据设定的特征维数对所述工业优选特征数据进行降维处理得到工业表征数据;

聚类模块,用于根据聚类模型、所述工业表征数据得到工业数据聚类结果;

所述优选特征确定模块包括:

特征全集获取子模块,用于获取所述预处理后的工业数据的所有特征得到特征全集;

第一轮特征筛选子模块,用于对所述特征全集进行第一轮特征筛选得到第一特征集合,所述第一轮特征筛选包括判断特征全集中每一特征是否唯一且有效,若是,则保留相应特征,若否,则删除或合并相应特征;

第二轮特征筛选子模块,用于对所述第一特征全集进行第二轮特征筛选得到第二特征集合,所述第二轮特征筛选包括判断第一特征集合中每一特征是否与产品性质密切相关,若是,则保留相应特征,若否,则删除相应特征;

工业优选特征数据获取子模块,用于从所述预处理后的工业数据中获取与所述第二特征集合中的特征对应的数据作为工业优选特征数据;

所述第二特征筛选子模块包括:

相关性计算单元,用于使用皮尔逊相关系数计算所述第一特征集合中每一特征与产品性质的相关性;

相关性阈值比较单元,用于判断所述相关性是否大于相关性阈值,若大于,则所述特征与产品性质密切相关。

一种工业数据处理方法和装置

技术领域

[0001] 本发明涉及数据处理技术领域,尤其涉及一种工业数据处理方法和装置。

背景技术

[0002] 石化生产过程的控制和产品收率的优化一直是石油加工领域研究的热点和难点,其中,石化生产过程中的工况信息对生产模式、以及产品类别和品质都具有重要的表征意义,可以使生产决策者和生产线操作人员正确了解当前的生产模式,并对产品收率的进一步优化起指导作用。

[0003] 然而在石化生产过程中,可收集的工况信息多达上千种,这就导致很难直观地展现某个具体的工况条件。

[0004] 于是,如何从多达上千种的工况信息种提取最重要的、与产品性质密切相关的特征,以更为直观的形式呈现出来,成为亟待解决的技术问题。

发明内容

[0005] 针对以上问题,本发明实施例提供了一种工业数据处理方法和装置。

[0006] 根据本发明实施例第一方面,一种工业数据处理方法,该方法包括:获取工业数据,工业数据包括传感器采集到的工况数据和原料数据;对工业数据进行预处理得到预处理后的工业数据;根据数据有效性和与产品性质的相关性,从预处理后的工业数据中确定工业优选特征数据;根据设定的特征维数对工业优选特征数据进行降维处理得到工业表征数据;根据聚类模型、工业表征数据得到工业数据聚类结果。

[0007] 根据本发明实施例一实施方式,根据数据有效性和与产品性质的相关性,从预处理后的工业数据中确定工业优选特征数据,包括:获取预处理后的工业数据的所有特征得到特征全集;对特征全集进行第一轮特征筛选得到第一特征集合,第一轮特征筛包括判断特征全集中每一特征是否唯一且有效,若是,则保留相应特征,若否,则删除或合并相应特征;对第一特征全集进行第二轮特征筛选得到第二特征集合,第二轮特征筛选包括判断第一特征集合中每一特征是否与产品性质密切相关,若是,则保留相应特征,若否,则删除相应特征;从预处理后的工业数据中获取与第二特征集合中的特征对应的数据作为工业优选特征数据。

[0008] 根据本发明实施例一实施方式,判断第一特征集合中每一特征是否与产品性质密切相关,包括:使用皮尔逊相关系数计算第一特征集合中每一特征与产品性质的相关性;判断相关性是否大于相关性阈值,若大于,则该特征与产品性质密切相关。

[0009] 根据本发明实施例一实施方式,在根据设定的特征维数对工业优选特征数据进行降维处理得到工业表征数据之前,该方法还包括:根据经验值设定特征维数,或根据降维效果设定特征维数。

[0010] 根据本发明实施例一实施方式,根据降维效果设定特征维数,包括:确定至少两个待定特征维数;计算使用每一待定特征维数进行特征降维后得到的数据方差;从至少两个

待定特征维数中选取数据方差最高的待定特征维数设定特征维数。

[0011] 根据本发明实施例一实施方式,在根据聚类模型、工业表征数据得到工业数据聚类结果之前,该方法还包括:根据聚类效果确定聚类算法和聚类模型;使用历史工业数据作为样本数据,对聚类模型进行训练和调优。

[0012] 根据本发明实施例一实施方式,根据聚类效果确定聚类算法和聚类模型,包括:选取至少两个待定聚类算法;根据聚类效果从至少两个待定聚类算法中确定一个聚类效果较好的待定聚类算法作为聚类算法;根据聚类算法建立聚类模型。

[0013] 根据本发明实施例一实施方式,根据聚类效果确定聚类算法和聚类模型,包括:使用k-means算法作为聚类算法;根据聚类效果确定k-means算法中要使用的簇集数量;根据k-means算法和簇集数量建立聚类模型。

[0014] 根据本发明实施例一实施方式,设定的特征维数小于等于3,相应地,在根据聚类模型、工业表征数据得到工业数据聚类结果之后,该方法还包括:以图形方式展示工业数据聚类结果以便为收率调优决策提供支持。

[0015] 根据本发明实施例第二方面,一种工业数据处理装置,装置包括:数据获取模块,用于获取工业数据,工业数据包括传感器采集到的工况数据和原料数据;数据预处理模块,用于对工业数据进行预处理得到预处理后的工业数据;优选特征确定模块,用于根据数据有效性和与产品性质的相关性,从预处理后的工业数据中确定工业优选特征数据;特征数据降维模块,用于根据设定的特征维数对工业优选特征数据进行降维处理得到工业表征数据;聚类模块,用于根据聚类模型、工业表征数据得到工业数据聚类结果。

[0016] 本发明实施例提供了一种工业数据处理方法和装置,该方法包括:首先,通过传感器实时采集工况数据并结合原料数据形成工业数据;之后,对工业数据进行预处理;然后,根据数据有效性和与产品性质的相关性,从预处理后的工业数据中确定工业优选特征数据;再根据设定的特征维数对工业优选特征数据进行降维处理得到工业表征数据;随后,利用聚类模型、工业表征数据得到工业数据聚类结果。如此,紧密结合产品性质对工业优选特征数据进行层层筛选得到更具代表性的工业表征数据,并根据这些表征数据和聚类模型得到的低维聚类结果,使工业数据代表的工况更具可视性,从而为收率寻优提供了重要参考依据。

[0017] 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

[0018] 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:

[0019] 在附图中,相同或对应的标号表示相同或对应的部分。

[0020] 图1为本发明实施例工业数据处理方法的实现流程示意图;

[0021] 图2为本发明实施例一应用使用聚类效果评价得分确定最佳簇集数量的示意图;

[0022] 图3为本发明实施例一应用将工业数据聚类结果进行图形展示的效果示意图;

[0023] 图4为本发明实施例工业数据处理装置的组成结构示意图。

具体实施方式

[0024] 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。

[0025] 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。

[0026] 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。

[0027] 根据本发明实施例第一方面，一种工业数据处理方法，如图1所示，该方法包括：操作110，获取工业数据，工业数据包括传感器采集到的工况数据和原料数据；操作120，对工业数据进行预处理得到预处理后的工业数据；操作130，根据数据有效性和与产品性质的相关性，从预处理后的工业数据中确定工业优选特征数据；操作140，根据设定的特征维数对工业优选特征数据进行降维处理得到工业表征数据；操作150，根据聚类模型、工业表征数据得到工业数据聚类结果。

[0028] 在操作110中，工况数据主要指传感器实时采集的、某一产品在生产过程中各个生产装置的运行状况、制备方法或生产条件等。原料数据主要指产生这些工况数据的生产过程中所采用的原料组成和原料配比等数据。原料数据通常是预先设定好的。这些数据基本上都是未加工的原始工业数据。

[0029] 对于有些制造工艺较为复杂的生产过程，有可能会产生上千种工况数据，且原始的工业数据大都是参差不齐的离散数据，并存在部分缺失、重复、噪音、异常等缺陷。因此，还需要进一步的清洗和再加工才能用来进行工况分析。

[0030] 在操作120中，对工业数据所进行的预处理主要指数据清洗和再加工，包括对存在的部分缺失、重复、噪音、异常等缺陷进行处理，以保证数据的完整性和正确性。

[0031] 举例来说，针对以下数据缺陷可以分别采取以下相应的处理方法：

[0032] 1)数据缺失：由于检测装置异常或者检测周期不一致的缘故，部分特征值或标签会存在缺失的情况，例如，某一时刻某一工业传感器或数据读写装置发生故障，导致某个小时内中有一段时间缺失该传感器所对应的数据。对于这种情况，可以根据缺失数据的重要程度来进行相应的处理。以石化工况数据为例，如果缺失的是和焦炭质量指标密切相关的关键数据，包括炉墙温度、炉内气压等，则可以剔除缺失这关键特征的整条工业数据；如果缺失的是焦炭质量指标无关的非关键数据，例如，燃炉标识等，则进一步判断缺失数值的数

量是否超出该条工况数据数值总数的三分之一,如果超出,则将该条工况数据判定为无效数据并进行剔除;如果没有超出,则可保留该条工况数据,并根据前后时间段的数据求均值,并使用均值补齐该条工况数据中的缺失数据。

[0033] 2)数据值异常:在传感器实时采集到的工况数据中,有时会因为设备故障出现一些数据值异常的数据,例如,在石化工况数据中的炉墙温度超出了正常温度的阈值,或在某个时间段温度激增等。首先,需要可以根据莱特准则,或与阈值比对的方法确定数据值异常的数据。在确定出数据值异常的数据之后,再根据这些数据的具体情况采用不同的措施。如果该条数据中的异常值不多,比如,小于三分之一,则可使用临近数据的该异常值对应的加权平均值替代该异常值;如果该条数据中的异常值较多,比如,大于等于三分之一,则可剔除整条数据。

[0034] 3)数据重复:有时在原始的工业数据中,由于种种原因可能会存在一些重复数据,例如,在石化工况数据中,发现时间点相同的两条炉墙温度数据。在这种情况下,可以对其中合法的数值求平均值,并以该平均值作为当前时间点的值。

[0035] 对工业数据进行预处理可以清除原始数据中的无效数据和异常数据,为后续的聚类分析提供清洁的数据基础。对数据的清洗越彻底,最后得到的工业数据聚类结果也越精准。

[0036] 在操作130中,工业特征是用以描述工业生产状态的一些指标,例如,各个生产装置的运行状况、制备方法、生产条件和原料配比等等。工业优选特征主要指经过筛选后的、有效的、且与产品性质密切相关的工业特征。此处的产品性质通常指决定产品品质的关键指标。例如,对于石化工业来讲,焦炭指标就是比较重要的一个产品性质。而工业优选特征数据则指与工业优选特征对应的数据。如前所述,对于一些制造工艺较为复杂的生产过程,有可能会产生上千种的工况数据,这些数据所对应的工业特征中难免会有一些意义重复的特征,或与产品性质无关的特征。保留与这些特征对应的数据,不但对获取工业数据聚类结果没有太大帮助,还可能还会增加数据计算的难度,甚至会对工况的评估和分析产生干扰。因此,本发明实施例工业数据处理方法还会对预处理后的工业数据进行进一步筛选,以得到有效的且与产品性质密切相关的工业优选特征数据。上述筛选主要是根据数据有效性和与产品性质的相关性来进行筛选的,例如,删除部分意义重复或无效的特征;根据行业经验进行筛选;根据采集指标与焦炭指标的相关性进行筛选等。

[0037] 上述过程中确定的工业优选特征数据是最终获得的工业数据聚类结果的数据基础,其选取的工业特征和产品性质越相关,获得的工业数据聚类结果对收率寻优就更具有指导意义。

[0038] 在操作140中,降维处理主要指在某些限定条件下,降低特征个数,得到一组“不相关”主特征的过程,其目的主要是为了进一步去除数据中的共线性特征,并移除异常样本,降低特征维度,使表征数据更为直观,易于描述。

[0039] 特征降维一般有两类方法:特征选择和特征抽取。特征选择即从高维度的特征中选择其中一个子集来作为新的特征;而特征抽取是指将高维度的特征经过某个函数映射至低纬度的特征作为新的特征。常用的特征抽取方法有主成分分析(Principal Components Analysis,PCA)等。PCA的实质就是在能尽可能保留原有特征的情况下,将原有特征进行线性变换、映射至低维度的空间中。

[0040] 此处的特征维数是预先设定好的,这一特征维数可以是根据经验值指定的,也可以是使用某一评估方法或工具计算或选择出来的。

[0041] 如前所述,对于某些制造工艺较为复杂的生产过程来说,即使经过特征筛选后的工业优选特征数据中的特征也是很庞大的,如果不进行降维处理,则很难从中得到直观、可视的工业数据聚类结果。所以这一操作也非常关键。

[0042] 在操作150中,聚类模型主要指用来进行聚类分析的模型,就是根据输入的数据进行聚类分析,输出聚类结果的模型。聚类分析指将物理或抽象对象的集合进行分组,由类似的对象组成的多个类的过程,聚类分析的目标就是在相似的基础上收集数据来进行分类。常用的聚类模型有K-mean聚类、层次(系统)聚类、最大期望EM算法等。

[0043] 在本发明实施例工业数据处理方法中,此处的工业数据聚类结果中的每一聚类通常对应于某种产品的制造工艺所对应的工况数据。

[0044] 通过对工业数据进行聚类,可以更好的筛选影响产品性质和收率的关键因素,为建立不同工况模式的收率预测模型提供数据基础,也为收率寻优提供寻优方向与指标参考。

[0045] 根据本发明实施例一实施方式,根据数据有效性和与产品性质的相关性,从预处理后的工业数据中确定工业优选特征数据,包括:获取预处理后的工业数据的所有特征得到特征全集;对特征全集进行第一轮特征筛选得到第一特征集合,第一轮特征筛包括判断特征全集中每一特征是否唯一且有效,若是,则保留相应特征,若否,则删除或合并相应特征;对第一特征全集进行第二轮特征筛选得到第二特征集合,第二轮特征筛选包括判断第一特征集合中每一特征是否与产品性质密切相关,若是,则保留相应特征,若否,则删除相应特征;从预处理后的工业数据中获取与第二特征集合中的特征对应的数据作为工业优选特征数据。

[0046] 在本实施方式中,主要采取先筛选特征,然后根据筛选后的特征来确定特征数据的方法来获取工业优选特征数据的。为了保证所选的工业特征是有效性的且和与产品性质密切相关,在本实施方式中,进行了两轮筛选。

[0047] 在本实施方式中,第一轮筛选的一个主要目的是删除一些无效的特征,例如,数据缺失高达百分之六十以上的特征,对于这些无效的特征,可直接删除;另一个主要目的则是合并意义重复的特征,比如生产装置运行状态中的环境温度和生产条件中的环境温度,虽然获取途径不同,但实际上代表的意义是相同的。在这种情况下,就可以对重复的特征进行合并。需要说明的是,这里的合并是保留其中的任一特征并删除其他特征。

[0048] 在本实施方式中,第二轮筛选主要是根据该工业特征是否与产品性质密切相关来进行筛选的。

[0049] 根据本发明实施例一实施方式,判断第一特征集合中每一特征是否与产品性质密切相关,包括:使用皮尔逊相关系数计算第一特征集合中每一特征与产品性质的相关性;判断相关性是否大于相关性阈值,若大于,则该特征与产品性质密切相关。

[0050] 在本实施方式中,皮尔逊相关系数又称皮尔逊积矩相关系数,是用于度量两个变量之间的线性相关,可以通过计算两个变量之间的协方差和标准差的商来得到,其值介于-1与1之间。通常情况下,可通过表1所示的取值范围判断两个变量的相关程度:

[0051]

皮尔逊相关系数的取值范围	两个变量的相关程度

0.8-1.0	极强相关
0.6-0.8	强相关
0.4-0.6	中等程度相关
0.2-0.4	弱相关
0.0-0.2	极弱相关或无相关

[0052] 表1

[0053] 在本实施方式中,就是通过样本计算每一特征与产品性质之间的协方差和标准差的商得到相关性,并与预设的相关性阈值进行比对来筛选与产品性质密切相关的特征。实施者可根据表1中给出的参考范围,以及想要达到的实施目标和实施效果来灵活制定筛选工业特征的相关性阈值。

[0054] 根据本发明实施例一实施方式,在根据设定的特征维数对工业优选特征数据进行降维处理得到工业表征数据之前,该方法还包括:根据经验值设定特征维数,或根据降维效果设定特征维数。

[0055] 根据本发明实施例一实施方式,根据降维效果设定特征维数,包括:确定至少两个待定特征维数;计算使用每一待定特征维数进行特征降维后得到的数据方差;从至少两个待定特征维数中选取数据方差最高的待定特征维数设定特征维数。

[0056] 通常,在特征降维后得到的数据方差越高,表示保留的特征信息越多。在本实施方式中,选取数据方差最高的待定特征维数来设定特征维数可以尽可能多地保留原有特征,使最后得到地工业数据聚类结果更准确。

[0057] 根据本发明实施例一实施方式,在根据聚类模型、工业表征数据得到工业数据聚类结果之前,该方法还包括:根据聚类效果确定聚类算法和聚类模型;使用历史工业数据作为样本数据,对聚类模型进行训练和调优。

[0058] 可以用于聚类效果评价的指标有很多,例如,轮廓系数、Calinski-Harabaz指数、兰德系数、互信息和V-measure等。实施者可以根据实际具体的实施条件和实施目标灵活选择。

[0059] 在进行训练的时候可以通过历史工业数据来训练聚类模型。历史工业数据和之后实际要聚类的数据通常是同一生产条件下、采用类似的制造工艺所得到的,与之后实际应用中实时采集到的工况数据更接近。如此,可以使得经过训练得到的模型能准确地对实时的工业表征数据进行聚类分析。使用这样的聚类模型得到的聚类结果也更精准。

[0060] 根据本发明实施例一实施方式,根据聚类效果确定聚类算法和聚类模型,包括:选取至少两个待定聚类算法;根据聚类效果从至少两个待定聚类算法中确定一个聚类效果较好的待定聚类算法作为聚类算法;根据聚类算法建立聚类模型。

[0061] 目前常用的聚类算法有K-Means聚类、均值漂移聚类、基于密度的聚类算法、层次聚类算法等等。在本实施方式中,可以先根据历史数据的特点初步选取至少两个算法,然后使用历史数据进行训练,并对每一算法训练得到的聚类结果进行聚类效果评价。之后,从中选取聚类效果评价较好的一个算法,并使用该算法建立聚类模型。

[0062] 根据本发明实施例一实施方式,根据聚类效果确定聚类算法和聚类模型,包括:使用k-means算法作为聚类算法;根据聚类效果确定k-means算法中要使用的簇集数量;根据k-means算法和簇集数量建立聚类模型。

[0063] 由于k-means算法是聚类分析中较为常用和通用的一种算法,只要确定一个合适的簇集数量,就可以在聚类分析的大多数应用场景中使用。在本实施方式中,并没有在几个聚类算法中进行选择,而是直接确定使用k-means算法作为聚类算法。

[0064] 关于簇集数量的确定,可以根据已有的经验值进行指定,该做法需要对生产数据本身以及生产状况有一定了解;也可以借助某种特定的方法,例如,聚类效果评价得分对比方法、或肘方法等,来估算一个最佳的簇集数量。

[0065] 图2就示出了本发明实施例一应用使用聚类效果评价得分确定最佳簇集数量的示意图。其中,横轴表示簇集的个数,纵轴表示划分成相应数量的簇集后的聚类效果评价得分。由图2可知,当簇集个数为2时,聚类效果评价得分最高。在这种情况下就可以将簇集个数指定为2。

[0066] 在指定了簇集个数之后,就可以根据k-means算法实现以下步骤来获取工业数据聚类结果:

[0067] 1)将工业表征数据转换为一个点,这个点可以是多维数字表示的一个向量;

[0068] 2)根据先验知识确定2个点做为初始聚集的簇心;

[0069] 3)分别计算每个点到每个簇心的距离(这里的距离一般取欧氏距离或余弦距离),找到离该点最近的簇心,并将该点划归到对应的簇;

[0070] 4)所有点都归属到簇之后,所有的工业表征数据就分为了2个簇。之后重新计算每个簇的重心(平均距离中心),并将该重心重新定义为新的簇心;

[0071] 反复迭代步骤3)到步骤4),直到达到某个中止条件,常用的中止条件有迭代次数小于预设的次数阈值、或最小平方误差MSE小于预设的平方误差阈值、簇中心点变化率小于预设的变化率阈值等。完成以上过程之后,就可以得到2个工业数据的聚类结果。

[0072] 根据本发明实施例一实施方式,设定的特征维数小于等于3,相应地,在根据聚类模型、工业表征数据得到工业数据聚类结果之后,该方法还包括:以图形方式展示工业数据聚类结果以便为收率调优决策提供支持。

[0073] 图3就示出了本发明实施例一应用将工业数据聚类结果进行图形展示的效果。在图3所示的应用中,指定的特征维数为3,簇集数为2,其中实心的点为一类簇集;空心的点为另一类簇集;每一簇集都代表某一生产工艺对应的工况。每个点都代表某一个小时的工况。在这一图中,x轴、y轴和z轴并没有具体的含义,就是每一工况数据经过特征降维后得到的三维工业表征数据的三个维度。此外,在该图中还可以显示某一工况最近的变化轨迹,即在指定时间段内,各状态点的时序连线,例如:图3所示的实线就表示一类簇集所代表的工况最近的变化轨迹;而图3中所示的虚线则表示另一类簇集所代表的工况最近的变化轨迹。此外,在实际的图形展示中还可以使用不同颜色的点代表最佳工况和最新工况,这些并未在图3中进行展示。

[0074] 在本实施方式中,通过图形展示工业数据聚类结果,可以更为直观地看到各种生产工艺所对应地工况情况,以及可以进行调优的方向等等,从而为收率寻优提供寻优方向与指标参考。

[0075] 根据本发明实施例第二方面,一种工业数据处理装置,如图4所示,该装置40包括:数据获取模块401,用于获取工业数据,工业数据包括传感器采集到的工况数据和原料数据;数据预处理模块402,用于对工业数据进行预处理得到预处理后的工业数据;优选特征

确定模块403,用于根据数据有效性和与产品性质的相关性,从预处理后的工业数据中确定工业优选特征数据;特征数据降维模块404,用于根据设定的特征维数对工业优选特征数据进行降维处理得到工业表征数据;聚类模块405,用于根据聚类模型、工业表征数据得到工业数据聚类结果。

[0076] 根据本发明实施例一实施方式,优选特征确定模块403包括:特征全集获取子模块,获取预处理后的工业数据的所有特征得到特征全集;第一轮特征筛选子模块,用于对特征全集进行第一轮特征筛选得到第一特征集合,第一轮特征筛包括判断特征全集中每一特征是否唯一且有效,若是,则保留相应特征,若否,则删除或合并相应特征;第二轮特征筛选子模块,用于对第一特征全集进行第二轮特征筛选得到第二特征集合,第二轮特征筛选包括判断第一特征集合中每一特征是否与产品性质密切相关,若是,则保留相应特征,若否,则删除相应特征;工业优选特征数据获取子模块,用于从预处理后的工业数据中获取与第二特征集合中的特征对应的数据作为工业优选特征数据。

[0077] 根据本发明实施例一实施方式,第二轮特征筛选子模块包括:相关性计算单元,用于使用皮尔逊相关系数计算第一特征集合中每一特征与产品性质的相关性;相关性阈值比较单元,用于判断相关性是否大于相关性阈值,若大于,则该特征与产品性质密切相关。

[0078] 根据本发明实施例一实施方式,该装置40还包括:特征维数设定模块,用于根据经验值设定特征维数,或根据降维效果设定特征维数。

[0079] 根据本发明实施例一实施方式,特征维数设定模块在具体用于根据降维效果设定特征维数时包括:待定特征维数确定子模块,用于确定至少两个待定特征维数;数据方差计算子模块,用于计算使用每一待定特征维数进行特征降维后得到的数据方差;特征维数设定子模块,用于从至少两个待定特征维数中选取数据方差最高的待定特征维数设定特征维数。

[0080] 根据本发明实施例一实施方式,该装置40还包括:聚类算法和聚类模型确定模块,用于根据聚类效果确定聚类算法和聚类模型;聚类模型训练模块,用于使用历史工业数据作为样本数据,对聚类模型进行训练和调优。

[0081] 根据本发明实施例一实施方式,聚类算法和聚类模型确定模块,包括:聚类算法筛选子模块,用于选取至少两个待定聚类算法;聚类算法确定子模块,用于根据聚类效果从至少两个待定聚类算法中确定一个聚类效果较好的待定聚类算法作为聚类算法;聚类模型建立子模块,用于根据聚类算法建立聚类模型。

[0082] 根据本发明实施例一实施方式,聚类算法和聚类模型确定模块,包括:聚类算法确定子模块,用于使用k-means算法作为聚类算法;簇集数量确定子模块,用于根据聚类效果确定k-means算法中要使用的簇集数量;聚类模型建立子模块,用于根据k-means算法和簇集数量建立聚类模型。

[0083] 根据本发明实施例一实施方式,该装置40还包括:图形展示模块,用于以图形方式展示工业数据聚类结果以便为收率调优决策提供支持。

[0084] 这里需要指出的是:以上针对工业数据处理的装置实施例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。对于工业数据处理的装置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。

[0085] 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。

[0086] 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。

[0087] 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。

[0088] 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。

[0089] 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。

[0090] 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。

[0091] 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。

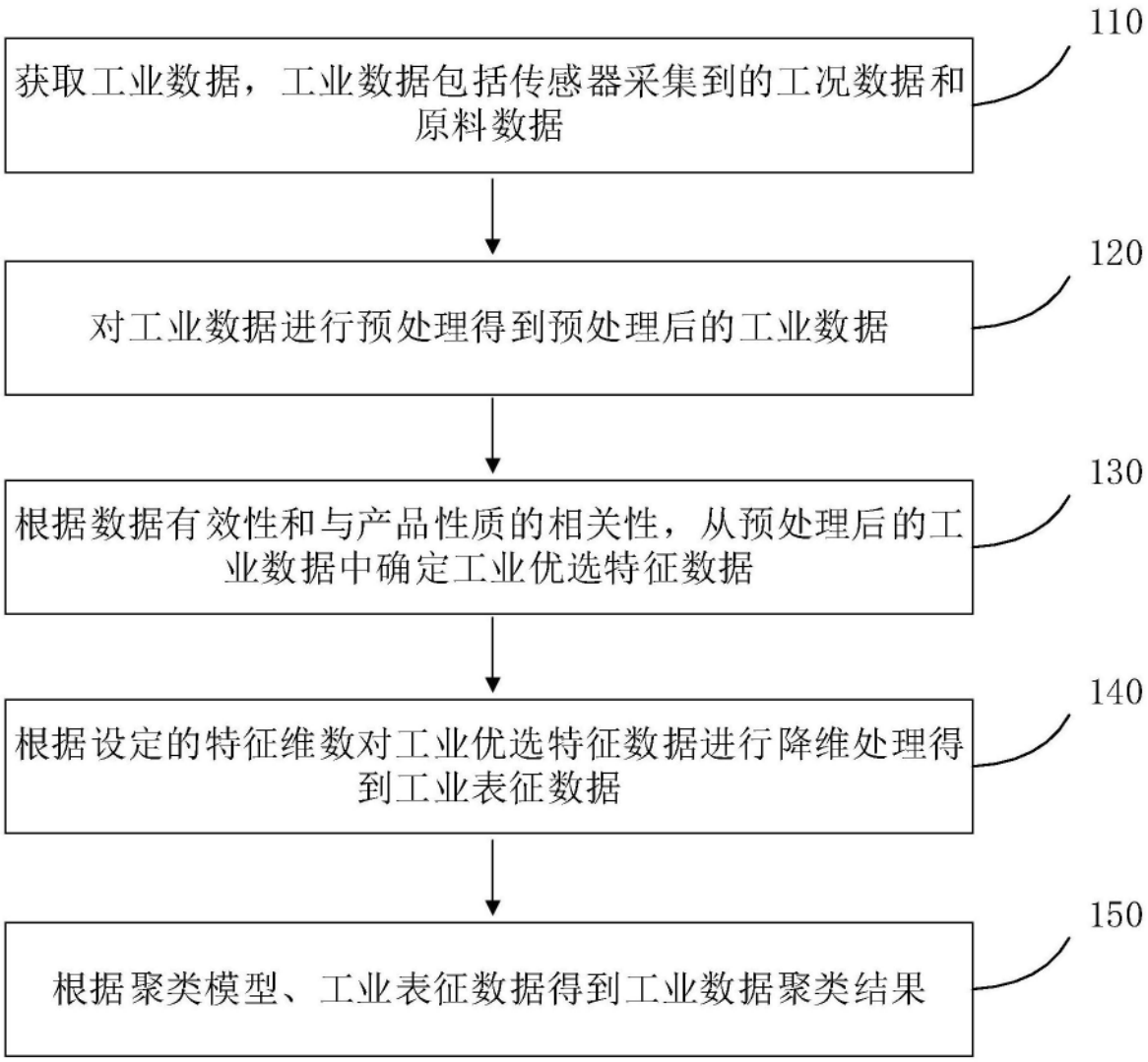

110
获取工业数据，工业数据包括传感器采集到的工况数据和原料数据
120
对工业数据进行预处理得到预处理后的工业数据
130
根据数据有效性和与产品性质的相关性，从预处理后的工业数据中确定工业优选特征数据
140
根据设定的特征维数对工业优选特征数据进行降维处理得到工业表征数据
150
根据聚类模型、工业表征数据得到工业数据聚类结果

图1

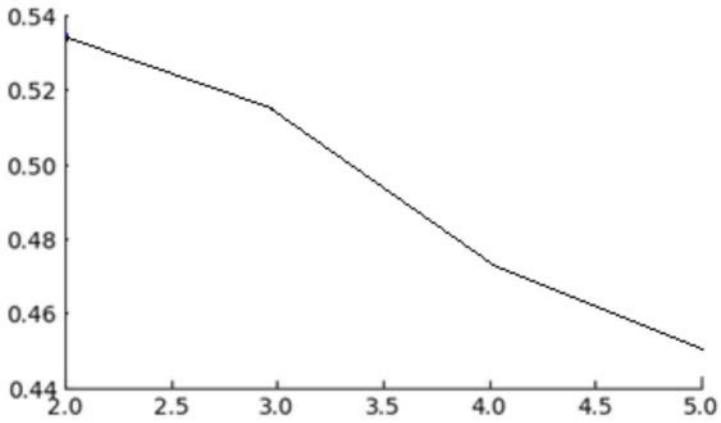

0.54
0.52
0.50
0.48
0.46
0.44
2.0
2.5
3.0
3.5
4.0
4.5
5.0

图2

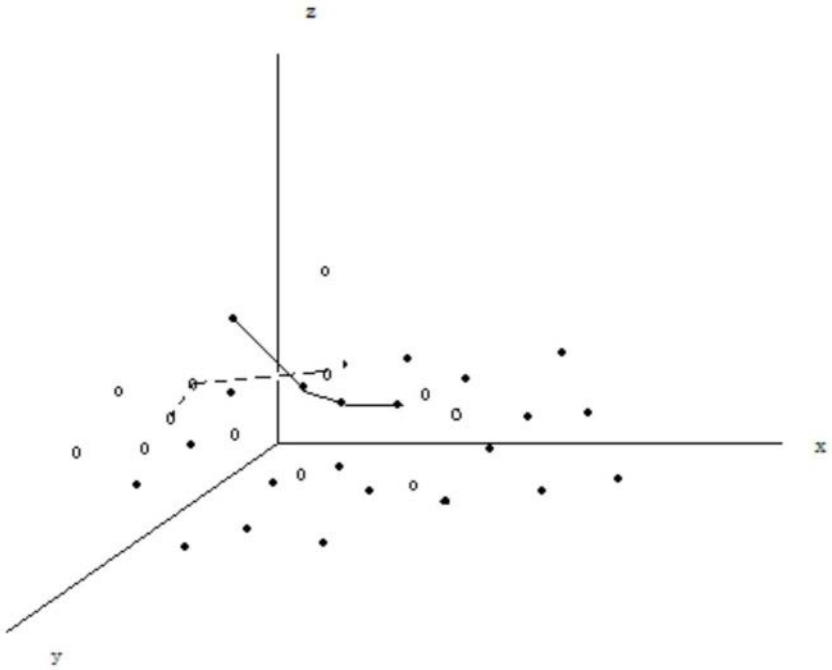

z
x
y

图3

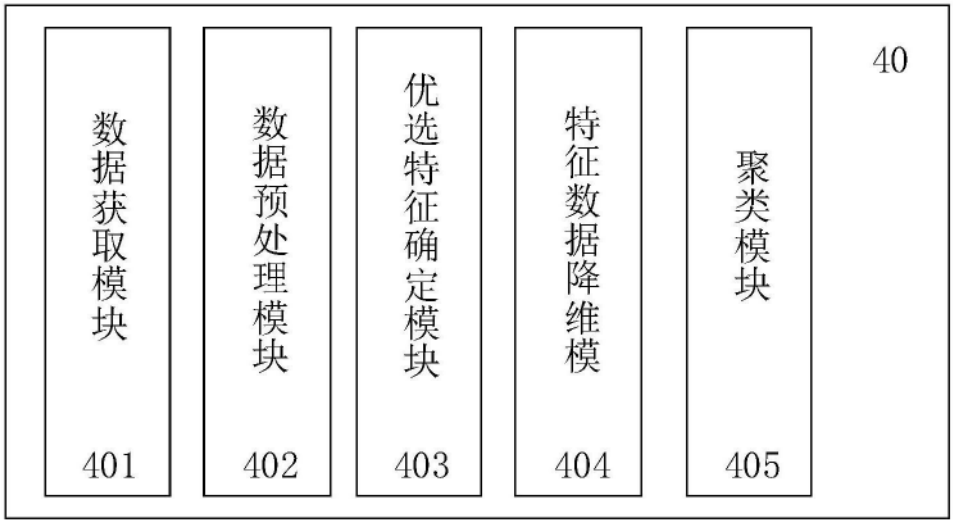

数据获取模块
401
数据预处理模块
402
优选特征确定模块
403
特征数据降维模
404
聚类模块
405
40

图4