基于云平台的压砖设备健康状态分析方法设计*

李晓昌[1],徐哲壮[1],谢仁栩[1],王 毅[1],刘 兴[1],王宏飞[1],夏玉雄[2]
(1.福州大学 电气工程与自动化学院,福建 福州 350108;
2.福建华鼎智造技术有限公司,福建 福州 350003)

摘 要:基于运行数据对压砖设备健康状态进行分析,对于降低设备故障率、提升压砖成品质量具有重要意义。现有方案大多数局限于离线人工分析,实时性差且推广效率低。针对上述问题,基于阿里云机器学习平台设计了压砖设备健康状态分析方法,基于聚类方法构建了压砖设备健康状态模型,在无需先验知识的情况下,对于压砖设备的工作、待机、异常等健康状态实现了建模。进而,将该模型部署于云计算平台上,通过周期性的数据导入与分析实现了压砖设备健康状态的在线分析。最后通过实例证明了该方法的有效性。

关键词:设备健康状态分析;工业大数据;机器学习;云平台;压砖设备
中图分类号:TP393 **文献标识码**:A DOI:10.19358/j.issn.2096-5133.2020.10.012
引用格式:李晓昌,徐哲壮,谢仁栩,等. 基于云平台的压砖设备健康状态分析方法设计[J].信息技术与网络安全,2020,39(10):61-66.

Design of health status analysis method for brick pressing machine based on cloud platform

Li Xiaochang[1], Xu Zhezhuang[1], Xie Renxu[1], Wang Yi[1], Liu Xing[1], Wang Hongfei[1], Xia Yuxiong[2]
(1.School of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, China;
2.Fujian Huading Intelligent Manufacturing Technology Co., Ltd., Fuzhou 350003, China)

Abstract: The analysis of the health status of the brick pressing machine based on the operating data is of great significance for reducing the failure rate of the machine and improving the quality of the finished brick press. Most existing solutions are limited to offline manual analysis, which has poor real-time performance and low promotion efficiency. In response to the above problems, this paper designed an analysis method of the health status of brick press machine based on the Alibaba Cloud machine learning platform. Based on the clustering method, the health state model of the brick press machine was constructed. Without prior knowledge, the health status of the brick press machine such as work, standby, and abnormality was modeled. Furthermore, the model was deployed on a cloud computing platform, and the online analysis of the health status of brick press machine was realized through periodic data import and analysis. An example was provided to prove the effectiveness of the proposed method.

Key words: machine health status analysis; industrial big data; machine learning; cloud platform; brick pressing machine

0 引言

工业设备的健康状态对于生产流程的稳定性与可靠性具有重要作用,单个设备故障会导致整条生产线停产,造成巨大的经济损失。因此,基于运行数据对工业设备健康状态进行分析,对于降低设备故障率、提升产品质量具有重要意义[1-3]。目前我国压砖产业已具备较大规模,新型压砖设备已能够通过工业物联网模块采集设备运行数据。但现有数据主要限于售后维护时使用,大量实时累计的运行数据并没有得到有效利用。另一方面,现有数据分析方案大多仍局限于离线人工分析,实时性差且推广效率低。因此,利用云平台[4-5]和机器学习技术[6-7]

*基金项目:国家自然科学基金资助项目(61973085);福州市市校科技合作项目(2018G86)

对设备健康状态进行在线分析已成为迫切需求[8]。

针对上述需求，本文基于阿里云机器学习平台设计了压砖设备健康状态分析方法，构建了压砖设备数据聚类分析模型，在无需专家先验知识的情况下，完成了压砖设备的工作、待机、异常等健康状态的建模。进一步地，通过将训练好的压砖设备健康状态模型部署至 DataWorks 平台，同时周期性地从保存压砖设备实时运行数据的 MySQL 数据库导出数据至该平台进行分析计算，实现了对压砖设备健康状态的在线分析。最后，本文通过实例证明了该方法的有效性。

1 压砖设备数据说明

压砖设备数据来自福建某压砖设备公司的设备监测平台，该平台通过工业物联网模块连接压砖设备生产线与本地服务器，实时采集压砖设备的数据至 MySQL 数据库中。每个制砖周期中都包含启动、振动压砖、待机等过程。在压砖过程中，需要通过两台振动电机的同步振动，才能保证压砖过程完成后砖块密度紧实。如果频繁出现振动电机振动不同步，则会导致电机故障、设备损坏、压砖成品均为无效等问题。

MySQL 数据库上存储的压砖设备数据中包含多个变量，涵盖了压砖设备振动电机 1 电流、振动电机 2 电流、油泵电流、三相电压等多个维度。本文分析所用数据的时间跨度从 2017 年 12 月初到 2018 年 10 月中旬，包含有近 600 万条压砖设备的运行状态记录。

2 基于云平台的压砖设备健康状态数据分析

本文采用阿里云机器学习平台 PAI(Platform of Artificial Intelligence)对压砖设备数据进行分析。压砖设备健康状态的数据分析流程主要由导入数据、数据预处理、特征分析、聚类分析和模型评估 5 个步骤组成，具体内容如下。

2.1 数据导入与预处理

在进入 PAI 平台并选择新建试验后，可以通过读数据表插件将离线数据加载到数据分析模块中。实时在线导入数据将在下文第 3.2 节详细介绍。

数据预处理主要分为删除缺失值、异常值处理、数据离散化、归一化处理等。在本文所获取的压砖设备数据中，存在数据的畸变值、缺失值等问题，会增加算法模型的复杂度，严重影响数据分析的精准性。本文采用了 PAI 平台的过滤与映射组件、缺失值填充组件，能够根据设置参数自动对数据集的成分进行筛选，处理掉缺失值与异常值。

2.2 特征分析

将数据导入平台的数据集中，包含有振动电机电流、三相电压、油泵电流等各种特征量。利用 PAI 平台的统计分析模块先对输入数据集中数据情况进行简单的统计分析，利用皮尔森系数分析各特征之间的相关性，在保证数据具有完整解释性的情况下，再对数据特征量进行筛选，有助于降低分析的复杂度。

(1)统计分析

数据预处理完成后，通过全表统计组件对压砖设备数据进行统计分析，得到了初始数据量为 1 437 148(行)×6(列)的矩阵。具体各变量统计区间如下：时间跨度(time)从 2017 年 12 月初到 2018 年 10 月中旬；振动电机 1 电流(i_1，单位 A)的范围为[0，38.7]；振动电机 2 电流(i_2，单位 A)的范围为[0，39.1]；油泵电流(oil_pump，单位 A)的范围为[0.1，114.4]；三相电压(voltage，单位 V)的范围为[2，816]。

(2)皮尔森相关性分析

通过给定压砖设备数据矩阵 X，计算 X 中两个特征列 i 和 j 的皮尔逊积矩(样本)相关系数的公式如式(1)所示：

$$R_{i,j}=\frac{\sum_{k=1}^{n}(X_{k,i}-\mu_{x_i})(X_{k,j}-\mu_{x_j})}{\sqrt{\sum_{k=1}^{n}(X_{k,i}-\mu_{x_i})^2\sum_{k=1}^{n}(X_{k,j}-\mu_{x_j})^2}} \tag{1}$$

其中，k 为变量，n 为 X 的总行数，μ_{x_i} 为 i 列数据的标准差，μ_{x_j} 为 j 列数据的标准差。$R_{i,j}$ 值的范围为[1，-1]。其中，1 表示具有强的正线性关系，-1 表示具有强的负线性相关，0 表示两变量之间没有线性关系。系数绝对值越趋于 1，相关性越大。

对压砖设备数据集 i_1、i_2、oil_pump、voltage 四个特征列进行相关性分析，得到如表 1 所示结果。由此可知，振动电机 1 电流变化情况与振动电机 2 电流变化情况基本一致，其相关系数约为 0.998，结合压砖设备的工艺特性可知，设备在制砖过程中需要

表 1 原始数据的皮尔森相关系数

特征量	电流 1	电流 2	三相电压
电流 1	1	0.997 9	0.300 3
电流 2	0.997 9	1	0.301 5
三相电压	0.300 3	0.301 5	1

两台振动电机的振动情况保持一致，这也是两者相关性大的主要原因。

根据$|R_{i,j}|>0.8$确定两数据量为强相关性的规则，结合砖机特征量的皮尔森相关系数表将线性相关性较强的数据特征量进行剔除，只保留i_1，但是又因为要通过振动电流的变化情况来判断电机的运行情况，因此本文通过特征之间的变换来创造特征量。

通过以上分析，最终确定了振动电机1电流(i_1)、电流差(i_dif)、三相电压(voltage)三个特征量来进行模型训练。

2.3 聚类分析

由于缺乏先验知识，在压砖设备上获取的数据是无标签的数据，因此本文采用K-means聚类算法对其进行分析处理[9-10]。PAI平台提供以多种距离的远近作为相似度测量值，包括euclidean距离、cosine距离、cityblock距离等，它通过不断迭代求得与初始聚类中心点的最优分类，使得评价指标达到预期。

本文采用欧式距离(euclidean距离)算法，计算各特征量与中心点之间的距离：

$$d(x-c)=(x-c)(x-c)' \tag{2}$$

设$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$为压砖设备数据样本集，其中包含$m$个未知运行状态的压砖设备数据特征。每个压砖数据样本$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{in})$是一个$n$维特征向量，则聚类样本集$D$划分为$k$个不相交的类别$\{C_l|l=1,2,\cdots,k\}$，其中$C_{l'\cap l'\neq l}C_l=\varnothing$且$D=\bigcup_{l=1}^{k}C_l$。相应地，用$\lambda_i\in\{1,2,\cdots,k\}$表示样本$\boldsymbol{x}_i$的“类标记”，即$\boldsymbol{x}_i\in C_{\lambda_i}$。于是聚类的结果可用包含$m$个元素的类标记向量$\boldsymbol{\lambda}=(\lambda_1;\lambda_2;\cdots;\lambda_m)$表示。聚类完成后可保存模型，加载至DataWorks中的算法节点实现在线数据分析，具体将在第3节进行介绍。

2.4 模型评估

本文采用Calinski-Harabasz(CH)指标作为聚类模型的评估指标。CH指标又称方差比率准则(Variance Ratio Criterion，VRC)，其定义为：

$$\mathrm{VRC}_k=\frac{\mathrm{SS}_B}{\mathrm{SS}_W}\times\frac{N-k}{k-1} \tag{3}$$

$$\mathrm{SS}_B=\sum_{i=1}^{k}n_i||m_i-m||^2 \tag{4}$$

$$\mathrm{SS}_W=\sum_{i=1}^{k}\sum_{x\in c_i}||x-m_i||^2 \tag{5}$$

其中，SS_B是整个聚类间的方差，SS_W是整个聚类内的方差，N是记录总数，k是聚类中心点个数，m_i是聚类i的中心点，m是输入数据的均值，x是数据点，c_i是第i个聚类。

由式(3)可得，聚类簇间区分度大，聚类簇内区分度小，因此可得VRC的值应越大越好。通过设定不同k值(聚类簇的数目)进行实验，得到了不同k值下的CH指标值，如图1所示。可以看到，不同k值下CH指标值变化波动性较大，不容易确定k的具体取值。因此将CH指标值较大对应k值(k=12，16，21)的特征中心点结果导出，并结合压砖设备的工艺特性与k值尽可能小的原则，最终选定k=16作为聚类数量。

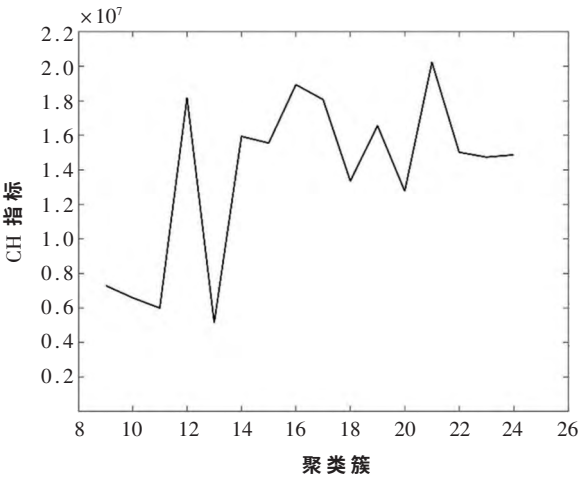

图1 CH值与聚类簇数变化情况

2.5 基于聚类结果的压砖设备健康状态分析

本节根据聚类算法所获得的压砖设备聚类簇中心点模型(表2)，结合压砖设备的工艺特征，对压砖设备的健康状态进行估计。

基于表2所示聚类簇中心点数据，可将设备健康状态归类和分析如下：

(1)严重异常：类别0、类别1、类别14聚类簇中心点的i_dif绝对值超过了10 A，两台电机的振动处于严重不同步的状态，将其定义为严重异常状态。在压砖过程中两台振动电机振动不同步，会影响压砖设备受力结构，并影响压砖成品质量。在这组数据中，此类数据总数很少，说明该设备仍处于正常状态。

(2)异常：类别2和类别13聚类簇中心点的i_dif绝对值小于10 A，但超过了2 A，将其定义为

表 2 压砖设备健康状态分析

类别	数量	电流 1/A	电流差/A	设备健康状态估计
0	4	3.95	-17.35	严重异常
1	46	6.18	-10.46	严重异常
2	1 652	16.50	-2.50	异常
3	993 622	0.06	-0.000 6	停机
4	230 509	4.93	-0.11	待机
5	28 931	6.95	-0.17	工况
6	42 610	9.54	-0.11	工况
7	84 808	14.28	-0.70	工况
8	23 537	15.79	-0.02	工况
9	14 726	17.31	0.05	工况
10	8 240	17.59	1.47	工况
11	3 904	20.86	-1.19	工况
12	2 713	19.43	1.89	工况
13	274	28.72	4.01	异常
14	38	17.4	10.02	严重异常
15	1 534	23.82	1.32	工况

异常状态。

(3)工况:类别 5、类别 6、类别 7、类别 8、类别 9、类别 10、类别 11、类别 12、类别 15 聚类簇中心点的 i_dif 绝对值不超过 2 A,i_1 的值大于 5 A,判断两台电机处于振动状态,且同时性较好,可定义其为工况状态。

(4)停机:类别 3 聚类簇中心点 i_1 和 i_dif 的值都几乎为 0,且总数很多,可定义其为停机状态。

(5)待机:类别 4 聚类簇中心点 i_dif 的值很小,但 i_1 在 5 A 左右,可定义其为待机状态。

3 压砖设备的在线健康状态分析

在通过聚类算法得到设备健康状态模型的基础上,可以进一步实现对压砖设备的在线健康状态分析。本文通过阿里云 DataWorks 平台,以设定的调度周期将实时采集至 MySQL 中的压砖设备数据同步至 MaxCompute 计算平台。DataWorks 平台与 PAI平台互通,将 PAI 平台上训练的模型部署至 DataWorks 平台,对在线数据进行健康状态分析,最后将分析结果导出至 MySQL,用于显示设备健康状态分析结果。完整工作流程如图 2 所示。

3.1 数据源

阿里云 DataWorks 支持多种数据源的接入,包括云数据库 RDS,对象存储 OSS、MySQL、Oracle 等。本文所用压砖设备数据存储在 MySQL 数据库中并且提供了公网可达的数据接口,在 DataWorks 工作空间的数据集成模块中添加 MySQL 数据源,通过 JDBC 连接串模式进行连接访问。本文将压砖设备数据的分析结果也存储在该 MySQL 数据库中。

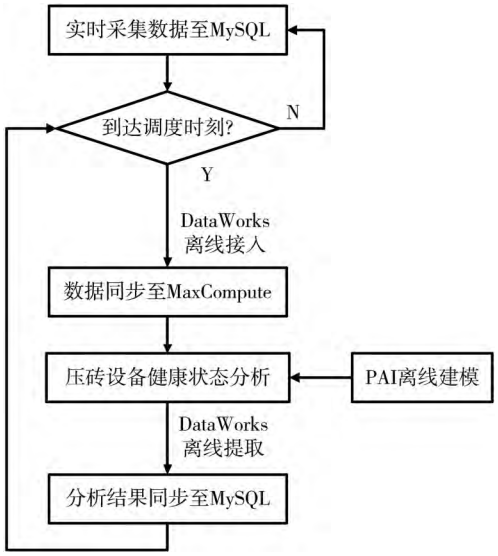

图 2 基于 DataWorks 平台的在线健康数据处理流程

3.2 数据同步节点

将数据从 MySQL 数据库接入 MaxCompute 计算平台有两种方式,即离线接入和实时接入,本文采取离线接入的方式将压砖设备数据接入 MaxCompute。MaxCompute 是一种快速、完全托管的 EB 级数据仓库解决方案,以数据为中心,内建多种计算模型和服务接口,可以满足广泛的数据分析需求[11]。

在 DataWorks 工作空间的数据开发模块中新建数据导入的数据同步节点,在 MySQL 数据库选择存储压砖机设备数据的数据表,将其与 MaxCompute 中接收接入数据的数据表对应起来,并将两个数据表的列逐一对应。设置数据接入时的过滤条件,模型训练完成后,过滤条件可根据调度周期来设定,确保新数据接入。

数据同步节点支持的调度周期包括分钟、小时、日、周和月。本文实验所采用的调度周期设置为 5 min。调度时间段可以根据实际工作时间设置,本文将其设置为全天24 h。配置好之后,每 5 min 内实时采集的压砖设备数据将从 MySQL 数据库导入 MaxCompute 进行分析计算。

3.3 数据算法节点

阿里云 DataWorks 平台与 PAI 平台互通,在 PAI 平台上训练模型,生成的模型可以发布到 DataWorks 使用,实现对机器学习实验的周期性调度。在

DataWorks 工作空间的数据开发模块中新建压砖设备健康状态分析算法节点，将训练的压砖设备数据模型离线部署至该算法节点，模型的训练过程在第 2 节已详细介绍。压砖设备健康状态分析算法节点的调度周期配置与数据接入节点的调度周期配置相同，确保及时对新接入的数据进行预测。

3.4 数据导出节点

本文将压砖设备的健康状态分析结果也存储在 MySQL 数据库中，在 DataWorks 工作空间的数据开发模块中新建分析结果导出数据同步节点，将分析结果从 MaxCompute 的数据表导入至 MySQL 的数据表中，该节点的调度周期配置与前两个节点一致，以确保预测结果及时导出。

需要说明的是，虽然数据导入节点、压砖设备健康状态分析算法节点、分析结果导出节点三者的调度周期都为 5 min，但是根据需求，三个节点并不是同时开始执行，而应该按数据导入→压砖设备健康状态分析→分析结果导出的顺序来依次执行。因此，将三个节点之间调度依赖的关系设置为：数据导入节点的输入为工作空间根节点，数据导入节点的输出为压砖设备健康状态分析算法节点的输入，压砖设备健康状态分析算法节点的输出为分析结果导出节点的输入。

3.5 压砖设备健康状态在线分析实例

基于上述方法，连续读取某压砖设备 1 h 内的运行数据进行健康状态分析，结果如图 3 所示。图 3 显示了振动电机 1 电流 i_1 和电流差 i_dif 的数据，同时每个数据所属的类别也进行了标记。可以看出压砖设备在前半小时不间断地进行周期性压砖，且基本处于正常工况状态。后半小时有部分时间段处于待机和停机状态。监测期间出现了 1 次严重异常记录，电流差 i_dif 值很大，但很快就恢复了正常值。运维人员可以根据严重异常记录的统计频次，判断设备的健康状态。在本实例中，此台压砖设备可判断为健康状态良好。

4 结论

本文基于阿里云机器学习平台设计了压砖设备健康状态分析方法，使用 K-means 聚类分析方法构建了压砖设备数据聚类分析模型，在无需专家先验知识的情况下，完成了压砖设备的工作、待机、异常等健康状态的建模。随后通过将训练好的压砖设备健康状态模型部署至 DataWorks 平台，同时周期性地从保存压砖设备实时运行数据的 MySQL 数据库导出数据至该平台进行分析计算，实现了对压砖

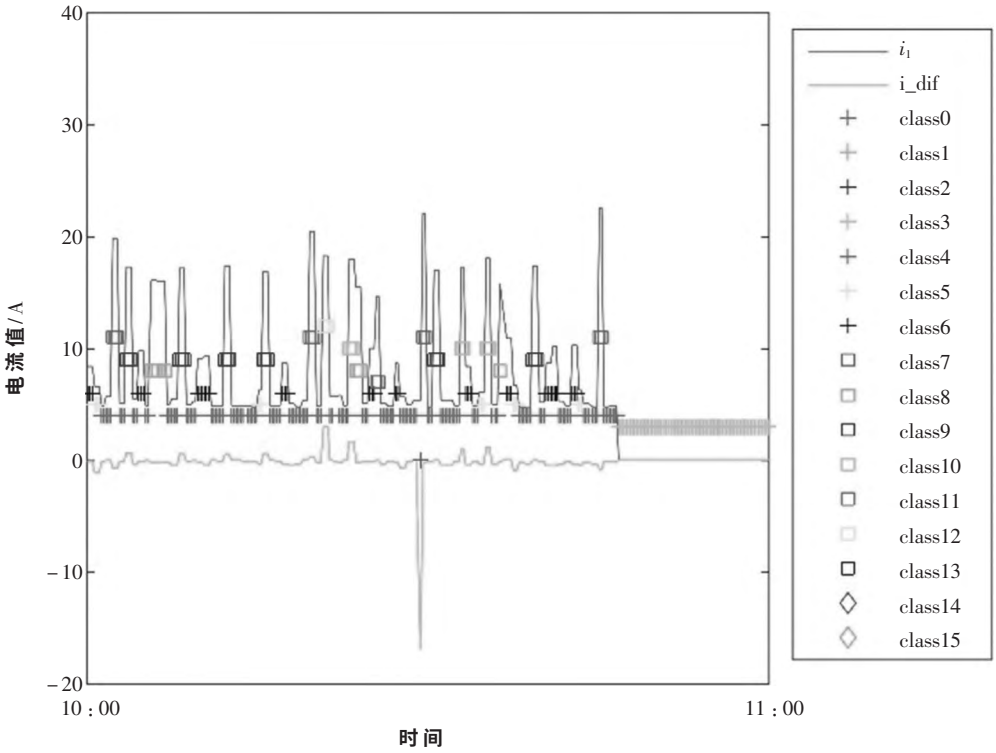

图 3 某压砖设备 1 h 内健康状态分析结果

设备健康状态的在线分析。最后，本文通过实例证明了方法的有效性。

参考文献

[1] 王海峰.基于工业大数据的智能化能耗管理与故障预诊系统[J].电气自动化，2018，40(4)：109－110.

[2] 冯言勇，刘桐杰，李昱，等.通过大数据分析进行设备故障诊断的技术研究[J].自动化博览，2017(12)：72－77.

[3] 谭晓鸿，贾磊朋.BP 神经网络故障诊断系统的仿真[J].硅谷，2010(12)：173－174.

[4] 牛禄青.阿里云：创新云计算[J].新经济导刊，2013(3)：66－68.

[5] 梅雅鑫.阿里云：打造三层边缘计算能力 构建云边端协同的开放生态[J].通信世界，2019(11)：44.

[6] 何清，李宁，罗文娟，等.大数据下的机器学习算法综述[J].模式识别与人工智能，2014，27(4)：327－336.

[7] 王桂玲，韩燕波，张仲妹.基于云计算的流数据集成与服务[J].计算机学报，2017，40(1)：107－125.

[8] 刘达新，裘乐淼，王志平.基于运行大数据学习的复杂装备故障诊断技术及其典型应用[J].中兴通讯技术，2017，23(4)：56－59.

[9] 章永来，周耀鉴.聚类算法综述[J].计算机应用，2019，39(7)：1869－1882.

[10] 周国亮，宋亚奇，王桂兰，等.状态监测大数据存储及聚类划分研究[J].电工技术学报，2013，28(S2)：337－344.

[11] 朱永利，李莉，宋亚奇，等.ODPS 平台下的电力设备监测大数据存储与并行处理方法[J].电工技术学报，2017，32(9)：199－210.

(收稿日期：2020－05－30)

作者简介：

李晓昌(1996－)，男，硕士研究生，主要研究方向：工业大数据。

徐哲壮(1984－)，通信作者，男，博士，副教授，主要研究方向：工业大数据、工业物联网与无线传感网等。E－mail：zzxu@fzu.edu.cn。

谢仁栩(1992－)，男，硕士研究生，主要研究方向：工业大数据。