文章编号：2095-2163(2023)02-0098-05 中图分类号：TF821;TP274 文献标志码：A

基于时序数据挖掘的铝电解槽工艺参数优化研究

张显国[1]，曹 斌[2]，王明刚[3]，石 进[1]

（1 贵州大学 大数据与信息工程学院，贵阳 550025；2 中铝智能科技发展有限公司 技术部，杭州 310000；
3 遵义铝业股份有限公司 分析计控中心，遵义 贵州 561300）

摘 要：中国铝电解生产过程通过获取在线监测数据和离线测量数据进行自动化控制，但是因工艺反应机理复杂、异构数据关联度低、作业控制大量依赖人工经验等问题，只能实现模糊控制，为了进一步提升电流效率、节能降耗，需要深入研究电解槽各项铝工艺参数间的数据相关度模型，本文基于铝电解槽生产时的时序数据，采用数据挖掘方法发现工艺参数之间关联性关系。首先，将预处理好的数据进行平稳性检验和正态分布检验，将满足性质的序列集合通过皮尔逊相关性分析、格兰杰因果检验和梯度提升回归树得到各个工艺参数之间的相关系数集合、因果关系集合和非线性关系系数集合，并建立工艺参数之间数学模型；其次，将非线性关系系数集合通过有向概率图建立物料平衡和热量平衡模型。以果变量出铝量为例，平衡模型检索出影响出铝量的所有工艺参数并进行验证，出铝量的拟合结果验证了本文方法的有效性。

关键词：数据挖掘；时序数据；铝工艺参数；因果关系；梯度提升回归树

Study on mechanism of aluminum electrolysis based on time series data mining

ZHANG Xianguo[1], CAO Bin[2], WANG Minggang[3], SHI Jin[1]

(1 College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China;
2 Technology Department, Chinalco Intelligent Technology Development Co., LTD., Hangzhou 310000, China;
3 Analysis and Control Center, Zunyi Aluminum Co., LTD., Zunyi Guizhou 563100, China)

【Abstract】 The production process of aluminum electrolysis in our country is automatically controlled by obtaining online monitoring data and offline measurement data. However, due to problems such as complex process reaction mechanism, low correlation degree of heterogeneous data, and a large amount of manual experience for operation control, fuzzy control can only be realized. In order to further improve current efficiency, energy saving and consumption reduction, an in-depth study of the data correlation model among the various aluminum process parameters of the electrolytic cell is required. Based on the time series data during the production of the aluminum electrolytic cell, this paper adopts the data mining method to discover the correlation relationship between the process parameters. Firstly, the preprocessed data is tested for stationarity and normal distribution, and the sequence sets satisfying the properties are analyzed by Pearson correlation, Granger causality test and gradient boosting regression tree to obtain the correlation coefficient between each process parameter set, causality set, and nonlinear relationship coefficient set, and establish a mathematical model between process parameters. Secondly, the nonlinear relationship coefficient set is used to establish a material balance and heat balance model through a directed probability graph. Taking the output variable of aluminum as an example, the balance model retrieves all the process parameters that affect the output of aluminum and verifies it. The simulation results of the output of aluminum verify the effectiveness of the method in this paper.

【Key words】 data mining; time series data; aluminum process parameters; cause and effect relationship; nonlinear regression relation

0 引 言

铝电解槽基本由 4 个部分组成：阴极结构、上部结构、母线结构以及电气绝缘部分。为保证电解槽安全稳定生产，需要工艺参数处于正常范围，如：电流、槽电压、极距、电解温度、电解质水平、加料次数等。由于电解铝生产环境的恶劣因素，如：强磁、高温、多粉尘和空间狭窄等，导致普通传感器不能在该环境下稳定运行，不能获取更多精准数据，如：分布电流、阴极温度和炉膛厚度等。目前能获取到的数据为单槽电流、单槽电压和一些原辅料下料量等间歇性测量和化验数据。

作者简介：张显国(1996-)，男，硕士研究生，主要研究方向：铝工业大数据时序处理技术；曹 斌(1963-)，男，博士，研究员，主要研究方向：机电一体化、工业控制、信息安全和智能管理。

通讯作者：曹 斌 Email：cb2027@126.com

收稿日期：2022-09-10

为提高铝液的产量和质量，挖掘物料、能耗等数据之间的关联性并建立铝电解反应模型具有重要意义。文献[1]提出使用数据挖掘技术得到平均电压、工作电压和效应持续时间之间的线性关系；文献[2]提出基于统计学方法控制电解槽的热量平衡；文献[3]提出基于贝叶斯网络的异常塑因模型。本文研究铝电解反应过程中的下料时序数据和生产工艺参数的时序数据，通过数据挖掘的方法，对铝电解槽内物料平衡和热量平衡建模，合成完整的铝电解槽反应过程模型。

1　相关工作

铝的电解化学反应式：

$$2Al_2O_3(\text{溶解})+3C(\text{固})\xrightarrow{\text{直流电}}4Al(\text{液})+3\ CO_2(\text{一次气体})$$

可知保持合适的物料平衡，一定范围内提高氧化铝浓度，可以提高铝的产量。

铝冶炼的出铝量 AL，公式(1)：

$$AL = 0.335\ 5 \times 24\bar{I}\gamma \sum (Nt) \times 10^{-6} \quad (1)$$

其中，0.335 5 为铝的电化学常数；$\bar{I}$ 为平均电流；γ 为电流效率；N 为电解槽数量；t 为运行昼夜数。

由公式(1)可知，在适宜的能量平衡下，提供电流值、电流效率或效应时长能够提高铝的产量。

通过数据挖掘方式获得物料平衡和能量平衡之间关联性，为防止后期数据挖掘出现虚假回归问题，先对所有序列进行平稳性检验。单位根检验(Augmented Dickey-Fuller test)是迪基-福勒检验(Dickey-Fuller test)的增广形式，其无漂移项回归公式如公式(2)所示：

$$\Delta X_t = \sigma X_{t-1} + \sum_{i=1}^{m} \beta_i \Delta X_{t-i} + \varepsilon_t \quad (2)$$

其中，Δ 为增量；ε_t 为 t 时刻残差(白噪声)；β_i 是 i 阶自回归加权系数 。

假设 H_0：$\delta = 0$，若检验序列存在单位根，则检验序列为非平稳序列，否则为平稳序列。

两个时间序列使用皮尔逊相关系数法需要满足以下条件：两个时间序列长度一致，连续且服从正态分布，因此首先检验时间序列是否服从正态分布，因为单维时序数据序列的样本数小于 5 000，所以采用夏皮罗-威尔克(Shapiro-Wilk)检验，根据检验结果检验序列是否服从正态分布。

夏皮罗-威尔克检验：单维时序数据序列是一个样本数为 n 的样本，假设 H_0：样本序列与正态分布没有显著区别，H_1：样本数据与正态分布有显著区别[4]。检验使用的统计量 W 定义为公式(3)：

$$W = \frac{\left(\sum a_i y_i\right)^2}{\sum (y_i - \bar{y})^2} \quad (3)$$

其中，y_i 代表真实值；$\bar{y}$ 是样本序列均值；$\left(\sum a_i y_i\right)^2$ 是 $(n-1)\sigma^2$ 的最佳线性无偏估计；σ 是来自样本的正态分布的标准差。

获得统计量后，设定显著性水平 α，获取其分位数或者临界值 W_α，若 $W < W_\alpha$，则检验序列符合正态性分布。

为降低格兰杰因果检验的计算复杂度，首先计算时序序列集合的相关系数集合。在指定时间段内，多维时序数据序列在时间段内 k 维时序数据序列之间的相关系数集合为 KR[5]，公式(4)：

$$KR = \begin{vmatrix} R_{11} & R_{12} & \cdots & R_{1k} \\ R_{21} & R_{22} & \cdots & R_{2k} \\ \vdots & \vdots & & \vdots \\ R_{k1} & R_{k2} & \cdots & R_{kk} \end{vmatrix} \quad (4)$$

其中，R_{ij} 表示 i，j 序列之间的相关系数值。

为了减少非线性关系分析的计算复杂度，先进行因果检验，获得序列之间的因果关系。格兰杰因果关系检验是一种推断和分析两个时间数据序列之间是否存在逻辑因果关系的检验算法[5]。检验序列 X 和检验序列 Y 在 T 时刻数值为 X_T 和 Y_T，公式(5)和公式(6)：

$$X_T = \sum_{i=1}^{T-1} \lambda_i X_i + \sum_{j=1}^{T-1} \delta_j Y_j + u_2 \quad (5)$$

$$Y_T = \sum_{i=1}^{T-1} \alpha_i X_i + \sum_{j=1}^{T-1} \beta_j Y_j + u_1 \quad (6)$$

其中，X_i 是序列 X 在 i 时刻的数值；Y_i 是序列 Y 在 i 时刻的数值；u_1 和 u_2 为不相关的白噪音；α，β，λ，δ 为参数。

若式(5)成立而式(6)不成立，则序列 Y 是引起 X 变化的因序列，存在序列 Y 到 X 的单向因果关系；若式(6)成立而式(5)不成立，则序列 X 是引起 Y 变化的因序列，存在序列 X 到 Y 的单向因果关系；若式(5)、式(6)同时成立，则认为 X 和 Y 存在双向因果关系。

为进一步理解因果变量之间的非线性程度，采用非线性回归分析方法，得到因变量和多个果变量的非线性回归系数集合。有助于关键参数的优化决策。梯度提升回归树有着较强的泛化能力，对异常值有很好的鲁棒性，以决策树为基函数，采用基函数

的线性组合与前向分布的提升方法,其基本思想是采用多个弱分类器构建一个强分类器[6]。

为获取影响出铝量的工艺影响参数路径图,基于图论和概率论以及贝叶斯网络的理论,构建有向概率无环网络图,其中节点表示铝电解槽的某个生产条件变量,有向边表示变量之间存在单向或者双向因果系。设有图 $G=(V,E)$,其中 $V=\{v \mid v \in S_k\}$,$E=\{e \mid e \in (R_{ij}=1)\}$,顶点 v 表示变量,路径 e 表示两个变量存在因果关系,箭头方向表示单向或者双向因果,节点概率值表示特征重要性程度[7]。

2 算法流程

首先,进行时序数据序列空值填充、重复值删除等预处理;其次,将满足正态分布的序列集合做皮尔逊相关处理,具有相关性的变量相互间进行格兰杰因果分析并得到因果变量集合,将满足平稳性的因果变量集合做非线性回归分析,得到因变量和多个果变量的非线性回归系数集合,对于不满足正态分布的变量、不满足平稳性的变量和其他没有相关性的变量在图中用孤立节点表示,最后将节点和工艺参数名对应并输出铝槽模型。

算法的具体流程如图 1 所示。

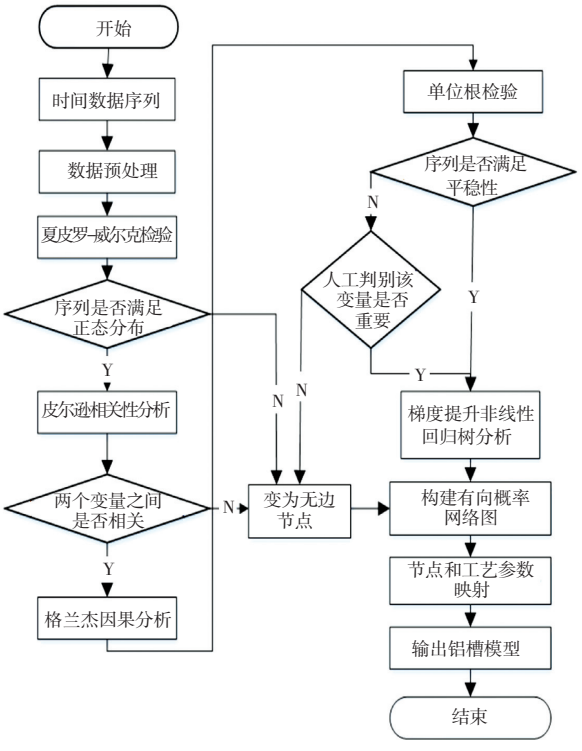

图 1 算法流程图

Fig. 1 Flow chart of algorithm

3 实验

3.1 数据处理

从铝厂的时序数据库中导出若干个铝电解槽的工艺参数的数据变量,包括日期、槽号、槽状态、运行时间、设定电压、工作电压、平均电压、效应电压、效应持续时间、效应次数、电压摆时间、异常持续时间、氧化铝下料次数、加料次数、氟盐添加次数、出铝指示量、基准下料间隔等。数据清洗方法如下:

由于传感器延迟传输导致的重复样本,本文根据时间戳保留第一个时间戳样本,删除其余重复样本;

若当前时间戳的工艺参数记录值缺失数量过多,则删除该样本,否则就采用众数填充的方式填补缺失值;

某些样本的某些属性值超出或者低于正常范围,为了保留真实的生产数据,不处理异常值并保留该样本。

3.2 挖掘序列性质

计算和获取单维时间序列的平稳性和正态性。采用单位根检验,检验的显著性结果 $p<0.05$,则该序列是平稳的时序数据序列。采用夏皮罗-威尔克方法检验每个时间序列,若统计量 W 小于 W_α,则检验序列符合正态性分布。根据检验结果的峰度、偏度以及图像形状判断序列是否满足正态分布,若序列峰度绝对值小于 10 和偏度绝对值小于 3,并且相应正态分布直方检验图呈现中间高,两边低的钟型,就判定检验序列符合正态分布。

3.3 挖掘序列关系

(1)相关关系。为获取多维时间序列之间的相关系数集合,将满足正态分布的数据集做皮尔逊相关系数处理。

(2)因果关系。为了判断两个工艺参数相互之间是否存在逻辑因果关系,选取具有相关性的时间序列进行格兰杰因果检验。实际设定电压和实际出铝量时间序列具有不平稳性,但在铝生产过程和指导出铝过程中具有重要意义,因此两个序列也要和其他序列做因果分析。

(3)非线性关系。使用梯度提升回归树算法计算影响果变量的各个因变量权值,数据集按照 8:2 划分训练集和测试集。梯度提升回归树节点分割的准则为弗里德曼均方误差[8],决策树的最大深度为 10,内部节点再划分所需最小样本数为 2,叶子节点最小样本数为 1,叶子节点样本最小权重为 0,最大

叶子节点数为 50。

根据每一组因果关系，梯度提升回归树获得非线性回归关系。

3.4　物料能量平衡模型

基于获得的因果关系集合和因果权值集合，构建贝叶斯网络结构。贝叶斯网络中的“节点”代表工艺参数，“有向边”代表两个工艺参数的因果关系，权值代表当前因变量影响果变量的程度。出铝关系图如图 2 所示。

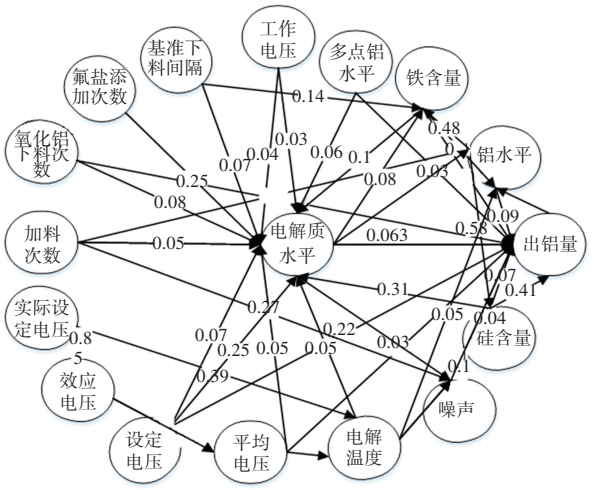

图 2　出铝关系图

Fig. 2　Aluminum output diagram

4　实验结果与结论

4.1　评价指标

为了定量分析模型对出铝量的拟合回归效果，采用可决系数 R^2、平均绝对误差（*MAE*）和均方根误差（*RMSE*）作为评价指标。

可决系数 R^2 值在[0,1]之间，数值越小代表模型越好，式(7)：

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

其中，$\bar{y}$ 代表平均值；y_i 代表真实值；$\hat{y}_i$ 代表模型预测值。

平均绝对误差（*MAE*），数值越小表示错误越小，模型越好，式(8)：

$$MAE(y,y) = \frac{1}{n_{samples}}\sum_{i=0}^{n_{samples}-1} | y_i - \hat{y}_i | \tag{8}$$

其中，$n_{samples}$ 代表样本数量；y_i 代表真实值；$\hat{y}_i$ 代表模型预测值。

均方根误差（*RMSE*）是在均方误差基础上求取平方根，式(9)：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{9}$$

其中，n 代表样本数量；y_i 代表真实值；$\hat{y}_i$ 代表模型预测值。

4.2　出铝量拟合结果

采用广度优先搜索算法，搜索出直接影响出铝量的直接节点，强相关关系系数阈值为 0.1，强因果关系的特征重要性程度阈值为 0.05。搜索出直接影响出铝量的因果关系，见表 1。得到氧化铝下料次数，加料次数、设定电压、针振、铝水平、电解质水平、铁含量、硅含量和出铝量的强相关系数绝对值在0.126和 0.331 之间，强因果关系的特征重要性程度在 0.07 和 0.22 之间。出铝因果关系表体现的物料平衡因果关系和公式(1)的化学反应方程表达的物料平衡结论基本一致。氧化铝下料次数和加料次数是氧化铝(Al_2O_3)和碳(C)的主要来源，频繁向电解槽添加适量的氧化铝，使得电解质中保持适当的氧化铝浓度和铝水平，提高铝的产量；电流和工作电压直接存在强相关关联。一定条件下，提高电压，从而提高电流，也能够提高铝的产量，和公式(2)表达的改变热量平衡来增加铝产量结论基本一致。

表 1　出铝量因果关系表

Tab. 1　Table of causality of aluminum output

配对样本		皮尔逊系数	特征系数
氧化铝下料次数		0.148	0.14
加料次数		0.233	0.16
设定电压	出	0.229	0.22
铝水平	铝	0.331	0.09
电解质水平	量	−0.154	0.06
铁含量		−0.263	0.09
硅含量		−0.249	0.07

根据出铝量的非线性关系集合，对包含氧化铝下料次数、加料次数、设定电压、工作电压、平均电压、噪声、铝水平、电解质水平、铁水平、硅水平，多点铝水平 11 个工艺参数的 14 组数据做非线性回归拟合分析。出铝量的非线性回归拟合结果见图 3，横坐标代表样本集编号，纵坐标代表出铝量。

由图 3 可知，氧化铝下料次数、加料次数、设定电压、铝水平、电解质水平、铁含量、硅含量能够影响出铝量，在已知因变量参数情况下能够预测出铝量。

采用评价指标分析出铝量的真实曲线和预测曲线拟合效果，得到可决系数、平均绝对误差和均方根误差，见表 2。由表 2 可知，该误差在铝工业生产的

可允许误差范围内,预测值 $\hat{y}_i$ 很接近真实值 y_i,拟合效果很好。

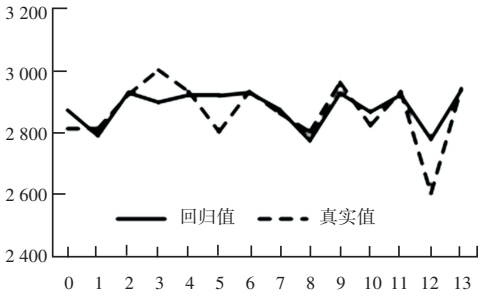

图 3 非线性回归拟合效果图

Fig. 3 Diagram of nonlinear regression fitting

表 2 评价指标表

Tab. 2 Table of evaluation indicators

评价指标	数值
*R*2	0.53
MAE	45.47
RMSE	67.57

5 结束语

在相关性分析和因果分析中,若出铝量是果变量,则有电解质温度、电解质水平、效应持续时间、硅含量、铁含量、铝水平、分子比、效应等待时间、各类电压、出铝指示量等 10 个因变量。在非线性回归分析中,10个因变量中有7个因变量和出铝量存在强因果关系,且因果权值在 7%~22%之间,影响权值总和为 83%。

本文基于多变量控制理念,通过对电解质温度、电解质水平、加料次数等工艺参数变量进行数据挖掘,得到各个工艺参数变量之间关联性,建立铝电解槽的物料平衡和热量平衡数学模型,实现铝电解槽对生产过程的参数优化和精确控制,达到提高电解效率和增加铝液产量的目的,对于推动铝电解槽增加出铝量具有重要意义。

参考文献

[1] 铁军, 朱旺喜, 吴智明. 数据挖掘技术在铝电解生产中的应用[J]. 有色金属, 2003 (1):56-59.

[2] 陈婷,康自华,曹斌. 基于统计过程控制方法的铝电解生产工艺优化[J]. 有色冶金设计与研究,2018,39(3):14-19.

[3] YUE W C, CHEN X F,GUI W H, et al. Aknowledge reasoning fuzzy - Bayesian network for root cause analysis of abnormal aluminum electrolysis cell condition [J]. Frontiers of Chemical Science and Engineering, 2017, 11(3):414-428

[4] 王沐贤,丁小欧,王宏志,等. 基于相关性的多维时序数据异常溯源方法[J]. 计算机科学与探索,2021,15(11):2142-2150.

[5] 冷喜武, 陈国平, 白静洁,等. 智能电网监控运行大数据分析系统总体设计[J]. 电力系统自动化, 2018, 42(12):7.

[6] 吕佳. 梯度提升回归树算法研究及改进[D]. 上海交通大学, 2017.

[7] 陈祖国, 李勇刚, 卢明,等. 基于贝叶斯概率语义网的铝电解槽况知识表示模型与约简方法[J]. 控制与决策, 2020, 35(7): 15.

[8] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

(上接第 97 页)

型有较好的识别效果,但因部分不同组分样本进行细化分类时,存在类别纹理特征相似情况,网络缺乏对这类组分细化辨识的能力,这种情况对网络识别准确率存在一定影响,后续研究可针对这一情况展开。

参考文献

[1] 陈浮,于昊辰,卞正富,等. 碳中和愿景下煤炭行业发展的危机与应对[J]. 煤炭学报,2021,46(6):1808-1820.

[2] LIU M, WANG P, CHEN S, et al. The classification of inertinite macerals in coal based on the multifractal spectrum method [J]. Applied Sciences, 2019, 9(24):5509.

[3] 王培珍,刘婕梅,汪文艳,等. 基于轮廓波变换的煤壳质组显微组分分类[J]. 煤炭学报, 2018, 43(S2): 641-645.

[4] 王培珍,殷子晓,王高,等. 一种基于 PCA 与 RBF-SVM 的煤岩显微组分镜质组分类方法[J]. 煤炭学报,2017,42(4): 977-984.

[5] 王培珍,刘曼,王高,等. 基于改进极限学习机的焦煤惰质组分类方法[J]. 煤炭学报,2020, 45(9): 3262-3268.

[6] 尹文花. 基于空间域数字图像处理方法的煤岩分析[D]. 太原:太原理工大学,2010.

[7] JI Z, XIA Y, SUN Q, et al. Fuzzy local Gaussian mixture model for brain MR image segmentation [J]. IEEE Transactions on Information Technology in Biomedicine, 2012, 16(3): 339-347.

[8] 朱楚雄,徐金明,钟传江. 基于全卷积神经网络的花岗岩中不同组分分布特征分析[J]. 中国质灾害与防治学报,2021,32(1): 127-134.

[9] 李荟,王梅. 用于大规模图像识别的特深卷积网络[J]. 计算机系统应用,2021,30(9):330-335.

[10] 孙海蓉,李号. 基于深度迁移学习的小样本光伏热斑识别方法[J]. 太阳能学报,2022,43(1):406-411.

[11] 王培珍,余晨,薛子邯,等. 基于迁移学习的煤岩壳质组显微组分识别模型[J]. 煤炭科学技术,2022,50(1):220-227.

[12] 陈方. MobileNet 压缩模型的研究与优化[D]. 武汉:华中师范大学,2018.

[13] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large - Scale Image Recognition [J]. Computer Science, 2015,4(6):8-22.

[14] WILSON M A, PUGMIRE R. J, et al. Carbon distribution in coals and coal macerals by cross polarization magic angle spinning carbon-13 nuclear magnetic resonance spectrometry[J]. Analytical Chemistry, 1984,56(6): 933-943.