doi：10. 12068/j. issn. 1005－3026. 2024. 03. 002

基于GMM-KNN-LSTM的烧结矿化学指标预测

闾光磊[1]，吴朝霞[1]，刘梦园[1]，姜玉山[2]
（1. 东北大学秦皇岛分校 控制工程学院，河北 秦皇岛 066004；
2. 东北大学秦皇岛分校 数学与统计学院，河北 秦皇岛 066004）

摘 要：针对烧结矿化学指标检测频率低导致无标签样本无法被机器学习利用的问题，提出了一种充分利用样本中有用信息的烧结矿化学指标预测模型. 首先，结合高斯混合模型（GMM）和K-近邻（KNN）算法，将无标签样本转化为有标签样本，然后与长短期记忆（LSTM）单元相结合，用于预测烧结矿的总铁质量分数、FeO质量分数和碱度3个化学指标. 通过与反向传播神经网络（BPNN）、循环神经网络（RNN）和LSTM三种模型对比，结果表明所建模型具有较低的预测误差. 总铁质量分数和FeO质量分数的预测命中率在允许误差±0. 5%内时分别达到98. 73%和95. 33%，碱度的预测命中率在允许误差±0. 05内为98. 13%，展现了较高的预测精度.

关 键 词：烧结矿化学指标；预测模型；无标签样本处理算法；LSTM；数据预处理
中图分类号：TF 046 **文献标志码**：A **文章编号**：1005-3026（2024）03-0314-09

Prediction of Sinter Chemical Indexes Based on GMM-KNN-LSTM

XIA Guang-lei[1], *WU Zhao-xia*[1], *LIU Meng-yuan*[1], *JIANG Yu-shan*[2]
（1. School of Control Engineering, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China; 2. School of Mathematics and Statistics, Northeastern University at Qinhuangdao, Qinhuangdao 066004, China. Corresponding author: WU Zhao-xia, E-mail: ysuwzx@126.com）

Abstract: Aiming at the problem that unlabeled samples cannot be utilized by machine learning due to the low detection frequency of sinter chemical indexes, a prediction model for sinter chemical indexes that makes full use of the useful information in the samples is proposed. Firstly, the unlabeled samples are transformed into labeled samples by combining Gaussian mixture model (GMM) and K-nearest neighbor (KNN) algorithm, and then combined with long short-term memory (LSTM) unit for predicting three chemical indexes, namely, total Fe mass fraction, FeO mass fraction, and alkalinity of sinter. By comparing with the three models of back propagation neural network (BPNN), recurrent neural network (RNN), and LSTM, the results show that the proposed model has a low prediction error. The prediction hit rates of total Fe mass fraction and FeO mass fraction reach 98. 73% and 95. 33%, respectively within the allowable error of ±0. 5%, and the prediction hit rate of alkalinity is 98. 13% within the allowable error of ±0. 05, demonstrating high prediction accuracy.

Key words: chemical indexes of sinter; prediction model; unlabeled samples processing algorithm; LSTM (long short-term memory); data preprocessing

烧结是钢铁工业的一个关键工序，它是高炉生产的重要原料. 烧结矿化学指标主要包括总铁含量、FeO含量和碱度，当烧结矿化学指标不合格时，会对高炉焦比和产量产生严重影响[1]. 因此，化学指标是衡量烧结矿质量的重要指标. 此外，烧结过程具有大滞后的特性，一个烧结矿的形成需要经历比例配料、加水混合和点火烧结3道工序，这大概需要1 h；然后烧结矿还需要经历破

收稿日期：2022-11-07
基金项目：河北省教育厅科学技术研究项目（BJ2021099）.
作者简介：闾光磊（1998-），男，安徽宣城人，东北大学硕士研究生；吴朝霞（1969-），女，河北秦皇岛人，东北大学教授.

碎、冷却和筛分才能得到成品烧结矿,这3个过程又需要大约1 h;而烧结矿化学指标的检测需要在实验室里花费更多的时间,最终的检测结果不能及时反映当前烧结过程的质量. 因为这种大滞后的存在,为烧结矿化学指标建立一个预测模型是十分必要的.

最近,随着工业人工智能技术的发展,许多研究已经将神经网络等机器学习算法成功应用到了钢铁行业过程参数的预测中[2-8],这些研究说明了机器学习算法应用于钢铁行业的可行性. 针对烧结矿化学指标的预测,国内外已经取得了一些研究进展. Li等[9]通过提出一种在线顺序极限学习机(OS-ELM)对烧结矿的FeO含量和转鼓指数进行预测,对烧结操作参数进行了优化,降低每吨烧结矿约0. 5 kg的燃料消耗. 吕庆等[10]采用XGBoost算法建立FeO含量预测模型,以指导生产工作人员及时调整配料方案和设备参数. 刘俊杰等[11]根据烧结矿化学成分与烧结工艺的预报、控制特点,采用了BP神经网络方法建立了烧结矿化学成分的预报模型. 仿真实验的结果表明,模型具有较高的预测精度和较强的自学习功能. 当前对烧结矿化学指标预测模型的研究虽然取得了一定的进展,但是由于烧结矿的化学指标数据只能在实验室化验得到,化学指标的化验结果通常需要4 h,这将导致产生的化学指标样本数据较少,而烧结过程的传感器可以采集大量的烧结过程数据,一般的策略就是将烧结过程样本数据的采集时间和化学指标样本数据的采集时间对齐,由于烧结过程样本数据远远多于化学指标数据,所以多出来的数据就变成了无标签样本,无标签样本是无法参与机器学习建模的,但是这些无标签样本含有大量有用的烧结过程信息,直接丢弃会对预测模型的准确性产生一定的影响. 所以当前研究普遍存在样本不够多或者烧结数据没有被充分利用的问题.

为了通过解决无标签样本的利用问题来提升预测模型精度,本文先用GMM-KNN算法把无标签样本变为有标签样本,GMM算法是一种聚类算法,可以很好地将烧结过程中具有相似特征的样本归为一类,而KNN回归算法可以从具有相似特征的有标签样本中找出和无标签样本最相似的样本,两者相结合可以将最相似样本的标签作为无标签样本的标签. 在解决无标签样本不能充分利用的问题后,采用LSTM算法来预测烧结矿的总铁质量分数、FeO质量分数和碱度,LSTM算法是一种循环神经网络,具有神经网络处理非线性系统的能力,同时可以学习时间序列的长短期相关信息[12],非常适合具有非线性、强耦合和时序性的烧结过程系统. 最后,本文基于烧结厂的实际生产数据,充分利用烧结过程无标签样本的有用信息,建立了基于GMM-KNN-LSTM的烧结矿化学指标预测模型,通过与其他几种常用神经网络算法对比,并对预测误差进行分析,从而验证本文所提模型的准确性和有效性.

1　数据描述及预处理

1.1　数据描述

烧结过程参数可以分为原料、混合料、操作、状态以及烧结矿化学指标,烧结过程主要参数如表1所示.

表1　烧结过程主要参数

Table 1　Main parameters of sinter process

参数	序号	参数名称	单位
原料	1	石灰粉	t/h
	2	除尘矿	t/h
	3	燃料	t/h
	4	铁粉	t/h
	5	烧结返矿	t/h
	6	高炉返矿	t/h
混合料	7	总铁质量分数	%
	8	五氧化二钒质量分数	%
	9	氧化钙质量分数	%
	10	二氧化硅质量分数	%
	11	水分质量分数	%
操作	12	圆辊转速	r/h
	13	九辊转速	r/h
	14	烧结机速度	m/min
	15	点火温度	°C
	16	煤气流量	m^3/h
	17	风机风量	m^3/h
状态	18	南烟道温度	°C
	19	南烟道负压	kPa
	20	北烟道温度	°C
	21	北烟道负压	kPa
	22~35	风箱废气温度	°C
	36~49	风箱负压	kPa
化学指标	50	烧结矿总铁质量分数	%
	51	烧结矿FeO质量分数	%
	52	烧结矿碱度	—

原料参数描述的是配料工序中料仓中原料的下料量;各个料仓的原料按照一定比例混合形成混合料,混合料参数就是混合料的化学成分,混合料经过布料装置进入烧结台车并进行点火烧结;操作参数是指点火烧结过程中可以调整的变量,以确保稳定地生产和更好的结果;状态参数描述的是烧结床的反应状态,烧结工人一般会通过这些状态参数判断烧结进展情况,进而对操作参数进行适当调节. 当烧结完成后,工人会在实验室里测量烧结矿的化学指标,但是化学指标的测定周期较长,这也是本文预测化学指标的主要原因.

1.2 缺失值和异常值处理

由于烧结过程生产环境干扰可能会导致传感器检测异常,造成部分数据的缺失和异常,这些缺失数据和异常数据对建立机器学习模型会产生严重影响,因此,在建立模型前需要对原始数据的缺失值和异常值进行处理. 识别异常值的经典方法有3σ原则和z得分法,使用这两种方法的前提假设是数据服从正态分布,但是实际烧结过程数据并不一定符合这个假设,而箱形图法是一种有效的异常值检测方法[13],它不需要事先假设数据服从特定形式的分布. 因此,本文采用箱形图法识别原始数据的异常值,箱形图如图1所示,其中,四分位($Q1$)、中位数(MD)和上四分位($Q3$)分别代表所有样本点值的25%, 50%和75%. 四分位距离(interquartile range, IQR)表示上四分位和下四分位之间的距离. 由于本文所用数据集的异常点不多,而神经网络等回归模型对异常值较为敏感,为了尽量减少异常值对神经网络的影响,选择温和异常点的边界$Q1-1.5$IQR和$Q3+1.5$IQR作为判断异常值的依据,即当数据点在箱形图的位置小于$Q1-1.5$IQR或大于$Q3+1.5$IQR,则该数据点为异常点. 筛选出来的缺失值和异常值需要进行处理,本文采用该数据点的前后5个正常数据点的平均值对它进行替换.

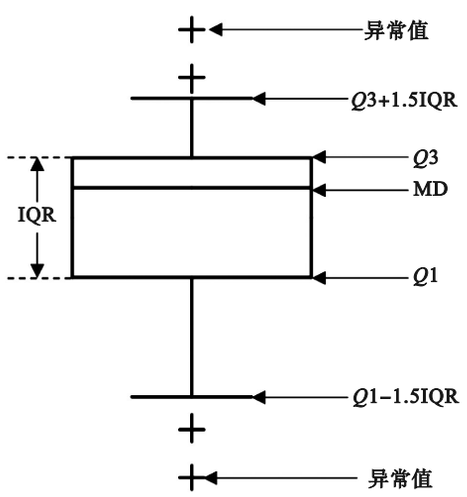

图1 箱形图

Fig. 1 Box plot

1.3 关键输入变量选择

由表1可知,烧结过程中存在高维过程变量,但是有些输入变量和预测化学指标无关,因此,本文首先从经验角度选出部分和输出变量相关的输入变量,在此基础上,采用统计学中的相关性分析法,在输入变量中筛选出和输出变量相关的变量. 变量之间的相关性可以分为线性和非线性,Pearson相关系数法可以计算出输入变量和输出变量的线性相关性,然而,对于具有强非线性的烧结过程来说,Pearson相关系数法并不是一个好的处理方法. 为此,本文引入可以计算非线性相关性的最大互信息法,提取与烧结矿总铁质量分数、FeO质量分数和碱度相关度最强的几个关键参数为输入变量,最大互信息系数MIC定义为

$$\mathrm{MIC}=\max_{n_1,n_2\leqslant B}\frac{\sum_{s}p_{x,y}(S)\mathrm{lb}\dfrac{p_{x,y}(S)}{p_x(S)p_y(S)}}{\mathrm{lb}(\min(n_1,n_2))}. \quad (1)$$

式中:n_1和n_2分别为x方向和y方向划分的区间数量;$p_x(S)$为样本点S落在x方向上的概率;$p_y(S)$为样本点S落在y方向上的概率;$p_{x,y}(S)$为样本点S落在x方向与样本点S落在y方向上的联合概率密度;B为划分的最大网格数量,为样本量的0.6次方.

2 烧结矿化学指标预测模型

GMM-KNN-LSTM模型预测化学指标的流程图如图2所示. GMM-KNN算法用于将无标签样本变为有标签样本,使LSTM算法可以利用更多的烧结过程信息,从而提升模型的预测精度. 将GMM-KNN算法和LSTM算法相结合建立烧结矿化学指标预测模型,用于预测烧结矿的总铁质量分数、FeO质量分数和碱度.

2.1 GMM-KNN无标签样本处理算法

由于烧结矿化学指标的样本数据量较小,导致烧结过程传感器数据只有一部分有标签样本,剩下的都是无标签样本. 烧结生产过程中,很多操作参数具有相似性,越相似的烧结生产操作越容易产生相近的烧结矿化学指标,虽然没有办法直接得到无标签样本对应的化学指标,但是可以把和无标签样本相似的样本标签作为它的标签. 本文通过将GMM算法和KNN算法相结合,先找

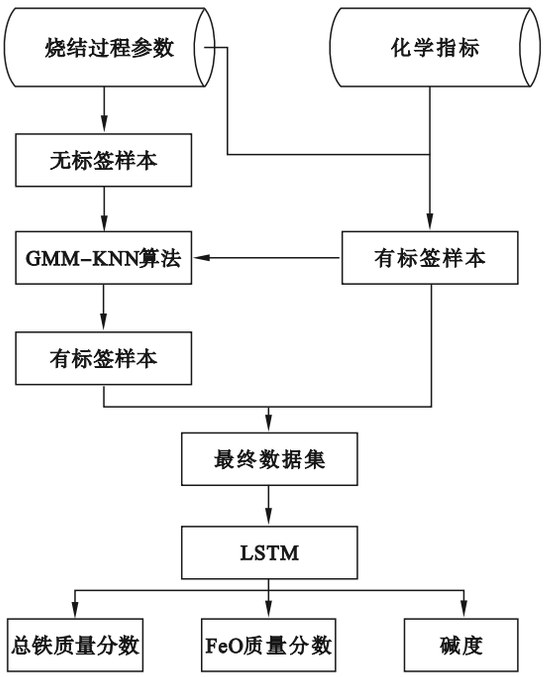

图2　GMM−KNN−LSTM模型预测流程图

Fig. 2　Flowchart of GMM-KNN-LSTM model prediction

到和当前无标签样本具有相似生产操作的有标签样本集合,然后在这个集合中找到最相似的5个有标签样本,再把这5个有标签样本对应的化学指标的平均值作为这个无标签样本的标签.

GMM算法用于根据概率分布将数据分类为不同的类别,烧结过程历史数据存在一些相似操作,GMM算法会将有标签数据集分为k个高斯分布,每一个高斯分布代表一个具有相似特征的类别,通过GMM算法把一些相似操作归为k类,然后可以计算某个无标签样本属于k类中的某一类.

定义有标签数据集$\boldsymbol{X}\{x_i\in\mathbf{R}^m,i=1,2,\cdots,N\}$,$m$为数据集的维度,它的高斯混合分布为

$$p(\boldsymbol{X}|\boldsymbol{\Theta}=[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_k])=\sum_{i=1}^{k}\alpha_i\cdot p(\boldsymbol{X}|\boldsymbol{\theta}_i). \quad (2)$$

式中:k是高斯成分的个数;$\boldsymbol{\Theta}=[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_k]$,$\boldsymbol{\theta}_i=[\mu_i,\Sigma_i]$,$\mu_i$是第$i$个高斯成分$C_i$的均值,$\Sigma_i$是第$i$个高斯成分$C_i$的协方差;$\alpha_i$为第$i$个高斯成分的混合系数,且$\sum_{i=1}^{k}\alpha_i=1$.

第i个高斯成分C_i的高斯密度函数为

$$p\left(\boldsymbol{X}|\boldsymbol{\theta}_i\right)=\frac{1}{\sqrt{(2\pi)^m|\Sigma_i|}}\mathrm{e}^{-\frac{1}{2}(\boldsymbol{X}-\mu_i)^{\mathrm{T}}\Sigma_i^{-1}(\boldsymbol{X}-\mu_i)}. \quad (3)$$

根据贝叶斯公式计算第j个样本x_j属于第i个高斯成分C_i的后验概率为

$$p_{ji}=\frac{\alpha_i p(x_j|\boldsymbol{\theta}_i)}{\sum_{i=1}^{k}\alpha_i p(x_j|\boldsymbol{\theta}_i)}. \quad (4)$$

通过式(4)分别计算某个无标签样本属于每个高斯成分的后验概率,最大后验概率对应的高斯成分就是这个无标签样本所属的相似类别n.

相似类别n里面有若干有标签样本,为了更精准地找到和无标签样本最相似的有标签样本,采用KNN算法从类别n中筛选出和对应无标签样本最相似的5个有标签样本.

KNN算法是工业中常用的机器学习算法[14],既可以用于分类也可以用于回归,当采用KNN回归算法时,可以找到某个样本的K个最相近的样本,并将它们标签的均值作为这个样本的标签. KNN算法是通过计算两个样本间的欧氏距离来判断它们的相似程度,距离越短,表示两个样本越接近. 对于两个样本x_i,x_j,它们之间的欧氏距离d可以表示为

$$d=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}. \quad (5)$$

式中:x_{ik}表示样本i的第k个特征值;x_{jk}表示样本j的第k个特征值.

烧结过程数据具有不同的量纲和单位,如果不作处理,会对欧氏距离计算结果产生很大影响,为了消除各维数据间数量级的差异,需要对数据进行标准化处理,即把所有数据都转化为0到1之间的数,标准化函数形式如下:

$$z=(x-\mu)/s. \quad (6)$$

式中:μ为样本均值;s为样本标准偏差.

2.2　LSTM算法

LSTM算法是RNN算法的一种变体,在处理时间序列时具有良好效果,广泛应用于具有时序特征的数据建模中[15]. LSTM算法的单元结构如图3所示.

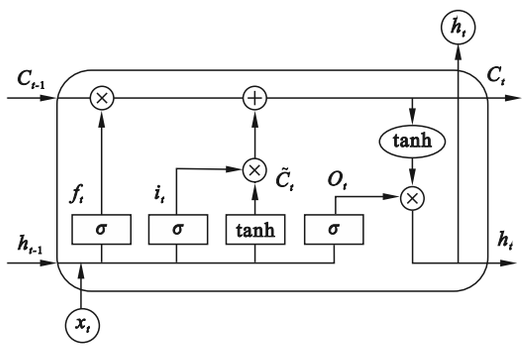

图3　LSTM算法结构图

Fig. 3　Structure diagram of LSTM algorithm

LSTM算法由1个记忆单元和3个门(遗忘门、输入门、输出门)组成,代替传统神经网络结构中隐藏层,使模型能够长期记忆,解决了传统循环神经网络梯度消失的问题.

遗忘门可以选择丢弃或者保留一些信息,激活函数是sigmod函数,输出在0到1之间,0表示完全舍弃,1表示完全保留. 遗忘门f_t表示为

$$f_t=\sigma(\boldsymbol{W}_{\mathrm{f}}h_{t-1}+\boldsymbol{U}_{\mathrm{f}}x_t+b_{\mathrm{f}}). \tag{7}$$

式中:σ为sigmod函数;$\boldsymbol{W}_{\mathrm{f}}$和$\boldsymbol{U}_{\mathrm{f}}$为遗忘门的权重系数;$h_{t-1}$为$t$−1时刻LSTM算法的输出;$x_t$为$t$时刻的输入;$b_{\mathrm{f}}$为遗忘门的偏置项.

输入门通过sigmod函数决定输入信息被保留的程度,输入门i_t表示为

$$i_t=\sigma(\boldsymbol{W}_{\mathrm{i}}h_{t-1}+\boldsymbol{U}_{\mathrm{i}}x_t+b_{\mathrm{i}}). \tag{8}$$

式中:$\boldsymbol{W}_{\mathrm{i}}$和$\boldsymbol{U}_{\mathrm{i}}$为输入门的权重系数;$b_{\mathrm{i}}$为输入门的偏置项.

LSTM算法状态的更新由临时记忆单元状态$\tilde{C}_t$和当前记忆单元状态C_t决定,它们分别表示为

$$\tilde{C}_t=\tanh(\boldsymbol{W}_{\mathrm{c}}h_{t-1}+\boldsymbol{U}_{\mathrm{c}}x_t+b_{\mathrm{c}}), \tag{9}$$

$$C_t=f_t*C_{t-1}+i_t*\tilde{C}_t. \tag{10}$$

式中:tanh为双曲正切函数;$\boldsymbol{W}_{\mathrm{c}}$和$\boldsymbol{U}_{\mathrm{c}}$为临时记忆单元的权重系数;$b_{\mathrm{c}}$为临时记忆单元的偏置项;$C_{t-1}$为$t$−1时刻记忆单元的状态;* 为Hadamard积.

输出门通过sigmod函数决定信息输出量,0表示完全不输出,1表示完全输出,输出门o_t可以表示为

$$o_t=\sigma(\boldsymbol{W}_{\mathrm{o}}h_{t-1}+\boldsymbol{U}_{\mathrm{o}}x_t+b_{\mathrm{o}}). \tag{11}$$

式中:$\boldsymbol{W}_{\mathrm{o}}$和$\boldsymbol{U}_{\mathrm{o}}$为输出门的权重系数;$b_{\mathrm{o}}$为输出门的偏置项.

LSTM算法最终的输出h_t可以表示为

$$h_t=o_t*\tanh(C_t). \tag{12}$$

2.3 模型评估指标

本文使用平均绝对误差(MAE)、均方误差(MSE)和均方根误差(RMSE)来评估点预测模型的准确度,误差越小,模型准确度越高. 它们的计算公式如下:

$$\mathrm{MAE}=\frac{1}{u}\sum_{i=1}^{u}|y_i-\hat{y}_i|, \tag{13}$$

$$\mathrm{MSE}=\frac{1}{u}\sum_{i=1}^{u}(\hat{y}_i-y_i)^2, \tag{14}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{u}\sum_{i=1}^{u}(y_i-\hat{y}_i)^2}. \tag{15}$$

式中:u为测试集的样本数量;$\hat{y}_i$表示第i个样本的预测值;y_i表示第i个样本的实际值.

3 结果分析与讨论

为了验证所提预测模型的准确性和有效性,从国内某烧结厂收集了真实的烧结过程数据,并对这些原始数据进行预处理,然后使用预处理后的数据建立本文所提预测模型,通过对比几种常见的神经网络算法来验证模型的有效性.

3.1 数据的异常值和缺失值处理结果

为了使输入数据可以达到建立模型的要求,需要对原始数据进行预处理. 缺失值是样本中少量为空的数据,可以直接由缺失值前后各5个样本点的均值代替. 异常值的识别采用箱形图法,箱形图法识别异常数据点的结果如图4所示,图4中的箱形图的顺序是按照表1中变量的顺序排列的. 由于各个参数存在量纲上的差异,将所

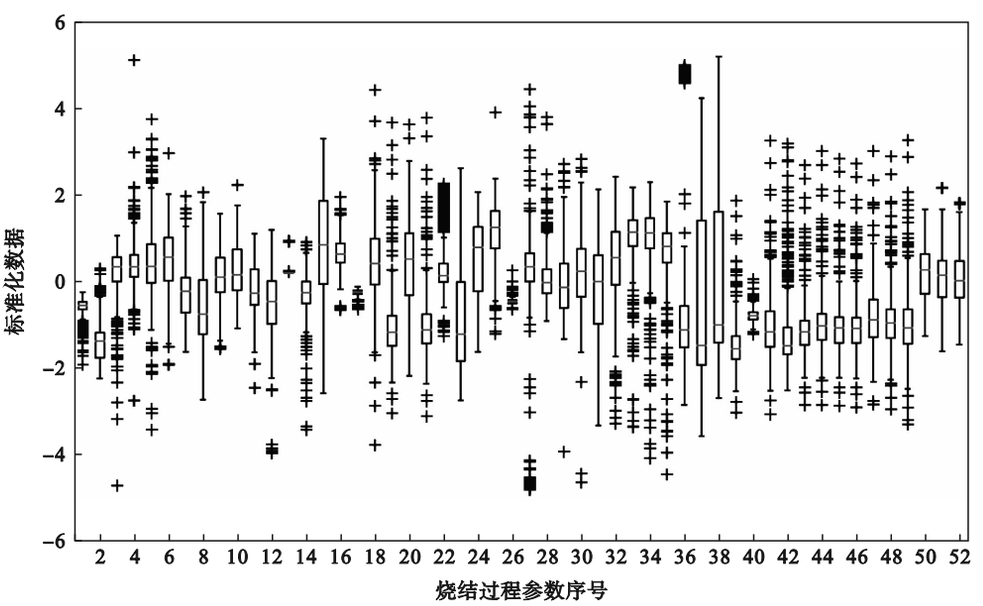

图4 箱形图法识别异常数据点的结果

Fig. 4 Detection results of outliers by box plot

有参数特征的箱形图画在一起会影响数据点的直观性,所以需要对所有特征数据进行标准化. 标准化是一种线性变换,同样的线性变换不会改变箱形图异常点的相对位置. 因此,标准化数据中的异常点的位置与原始数据的位置相同. 图4中共有52个箱形图,每个箱形图展示了对应烧结过程参数的样本值大小分布. 图4中黑色十字形状标记的点为该参数的异常点,可以分别找出52个箱形图对应参数的样本异常点,然后用异常点前后各5个正常样本点的均值来代替.

3.2　关键输入变量选择结果

分别计算49个输入参数与总铁质量分数、FeO质量分数和碱度3个化学指标的MIC值,取3个化学指标的MIC平均值,按照MIC平均值从大到小的顺序选出的32个输入变量见表2. 由表2可知,各个过程参数和3个化学指标的MIC值的相对大小顺序基本一致,这是因为各个化学反应之间是相互作用、相互影响的,所以和3个化学指标相关的参数非常相近,和化学指标相关的过程参数主要有风机风量、风箱负压、风箱废气温度和原料下料量. 风机风量和风箱负压是烧结过程中的重要参数,如果风机风量不足,将会直接影响烧结过程的好坏,风箱负压是风箱内的压力和外界大气压的差值,由于负压的作用,料层的燃烧层可以向下充分燃烧;风箱废气温度可以反映料层表面的温度变化,烧结工人一般会通过风箱废弃温度来判断烧结终点的位置,而烧结终点的位置直接影响烧结矿的质量. 铁粉、石灰粉、除尘矿、返矿和燃料是烧结矿的原料,烧结矿总铁质量分数受到所有含铁原料配比的影响,FeO质量分数会受到燃料配比和混合料碱度的影响,烧结矿碱度会受到石灰粉配比的影响,而各个化学成分之间是互相影响的,所以不同原料的配比对烧结矿化学指标有重要影响.

表2　烧结过程中3个化学指标的MIC值及其均值

Table 2　MIC values and their mean values of three chemical indexes in sintering process

过程参数	MIC值			MIC均值	过程参数	MIC值			MIC均值
	总铁质量分数	FeO质量分数	碱度			总铁质量分数	FeO质量分数	碱度	
风机风量	0.797 9	0.628 7	0.606 2	0.677 6	7号风箱废弃温度	0.681 9	0.512 8	0.490 7	0.561 8
7号风箱负压	0.602 5	0.494 3	0.458 3	0.518 4	铁粉	0.589 0	0.479 2	0.476 1	0.514 8
石灰粉	0.583 4	0.456 2	0.467 1	0.502 2	除尘矿	0.564 2	0.415 6	0.457 3	0.479 1
3号风箱废气温度	0.540 8	0.431 6	0.445 7	0.472 7	9号风箱负压	0.519 5	0.422 2	0.420 0	0.453 9
燃料	0.529 1	0.404 9	0.404 7	0.446 2	烧结返矿	0.464 4	0.402 4	0.375 2	0.414 0
22号风箱负压	0.482 8	0.356 3	0.376 4	0.405 2	11号风箱负压	0.466 8	0.354 4	0.388 6	0.403 3
3号风箱负压	0.479 9	0.343 6	0.372 9	0.398 8	15号风箱负压	0.476 6	0.348 5	0.370 0	0.398 4
13号风箱负压	0.473 7	0.344 8	0.366 7	0.395 1	21号风箱负压	0.476 6	0.336 6	0.368 6	0.393 9
5号风箱负压	0.466 7	0.351 2	0.359 8	0.392 6	20号风箱负压	0.470 5	0.339 1	0.367 5	0.392 4
16号风箱负压	0.467 8	0.342 0	0.363 7	0.391 2	18号风箱负压	0.468 9	0.337 2	0.364 4	0.390 2
南烟道负压	0.460 1	0.334 6	0.353 0	0.382 6	1号风箱负压	0.447 8	0.352 6	0.344 7	0.381 7
2号风箱负压	0.446 7	0.307 1	0.325 0	0.359 6	混合料水分	0.413 7	0.308 6	0.318 6	0.347 0
5号风箱废气温度	0.347 3	0.288 5	0.285 4	0.307 1	21号风箱废气温度	0.341 1	0.289 2	0.261 2	0.297 2
1号风箱废气温度	0.333 1	0.273 2	0.280 4	0.295 6	高炉返矿	0.288 8	0.295 6	0.282 3	0.288 9
20号风箱废气温度	0.310 4	0.271 8	0.268 7	0.283 6	22号风箱废气温度	0.307 1	0.262 7	0.265 2	0.278 4
北烟道温度	0.338 1	0.243 1	0.252 0	0.277 7	烧结机速度	0.333 0	0.250 4	0.248 8	0.277 4

3.3　无标签样本处理结果

本文收集的烧结数据包括39 522个烧结过程传感器采集的实时数据和13 174个烧结矿化学指标检测数据.

由于检测频率的不同,传感器采集的样本数量是化学指标检测到的样本数量的3倍. 在不对数据进行处理的情况下,39 522个传感器数据中只有13 174个样本能有对应的化学指标检测数据,也就是说,只有13 174个样本能参与建模,剩下的26 348个样本成为无标签样本数据,无法参与建模. 然而这些无标签样本数据中含有烧结过程中的大量有用信息,为了充分利用传感器采集的数据,使用GMM−KNN算法处理无标签数据,使无标签数据也可以参与建模,GMM−KNN算法处理样本的流程图如图5所示.

经过GMM−KNN算法处理后,使得无标签

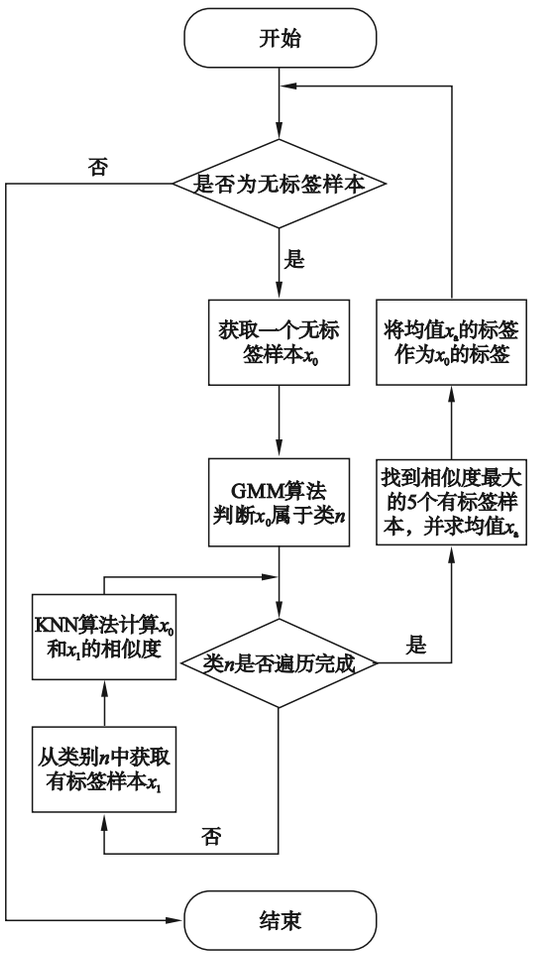

图5 GMM-KNN算法处理样本的流程图

Fig. 5 Flowchart of GMM-KNN algorithm processing samples

数据有了对应的标签，获得了总数39 522个有标签数据，减少了烧结过程有效数据的浪费，大大增加了可用于训练的标签样本数.

3.4 模型预测性能对比与分析

在得到关键输入变量和处理无标签样本之后，将输入样本按8∶2的比例分为训练集和测试集，然后使用训练集进行模型训练，最后用测试集去测试训练好的模型，采用评估公式对模型的性能进行评估，本文使用的软件环境为Python3. 8和scikit-learn1. 0. 2，硬件环境为型号AMD R7-4800h的CPU和容量64 GB的RAM. 采用GMM-KNN-LSTM模型预测烧结矿化学指标，为了更好地展示GMM-KNN-LSTM模型的性能，将它与DNN，RNN和未进行GMM算法处理的普通LSTM模型的预测性能进行对比. 由于运行时模型计算时间的长短决定模型是否可以快速得到烧结矿化学指标的预测结果，所以本文记录了不同模型在测试集上运行时的计算时间，BPNN，RNN，LSTM和GMM-KNN-LSTM模型的计算时间分别为0. 004，0. 096，0. 283 7和0. 283 7 s，由于LSTM模型的网络结构特征最复杂，导致其计算耗时最长，但是模型计算时间在1s内，可以快速得到烧结矿化学指标的预测结果. 4种不同模型的评估结果见表3～表5.

表3 不同模型的总铁质量分数预测性能比较

Table 3 Comparison of prediction performance of total Fe mass fraction with different models

模型	MSE	MAE	RMSE
BPNN	0. 083	0. 215 3	0. 287 7
RNN	0. 063	0. 186 7	0. 250 6
LSTM	0. 056	0. 169 1	0. 237 4
GMM-KNN-LSTM	0. 017 5	0. 089 1	0. 132 2

表4 不同模型的FeO质量分数预测性能比较

Table 4 Comparison of prediction performance of FeO mass fraction with different models

模型	MSE	MAE	RMSE
BPNN	0. 212 5	0. 346 3	0. 461 0
RNN	0. 161 7	0. 300 0	0. 402 1
LSTM	0. 148 9	0. 275 5	0. 385 8
GMM-KNN-LSTM	0. 034 7	0. 128 3	0. 186 4

表5 不同模型的碱度预测性能比较

Table 5 Comparison of prediction performance of alkalinity with different models

模型	MSE	MAE	RMSE
BPNN	0. 001 1	0. 024 6	0. 032 5
RNN	0. 000 8	0. 021 4	0. 028 7
LSTM	0. 000 7	0. 019 7	0. 027 7
GMM-KNN-LSTM	0. 000 1	0. 009 0	0. 013 6

由表3~表5可看出，GMM-KNN-LSTM模型的MSE，MAE和RMSE都小于其他3种模型，因为GMM-KNN-LSTM模型充分利用了大量有用的无标签样本信息，使得预测结果更精确. 除此之外，BPNN模型由于没有考虑时间维度的有用信息，所以精度低于循环神经网络，而LSTM模型是对RNN模型的一种改进，它很好地改善了RNN模型的梯度消失问题，所以LSTM模型的预测误差低于DNN模型和RNN模型.

MSE，MAE和RMSE都是衡量模型的平均性能，平均性能不能反映局部的预测情况，为了展示GMM-KNN-LSTM模型预测结果的可靠性，需要统计预测误差的分布情况，总铁质量分

数、FeO质量分数和碱度的预测误差频数直方图如图6所示.

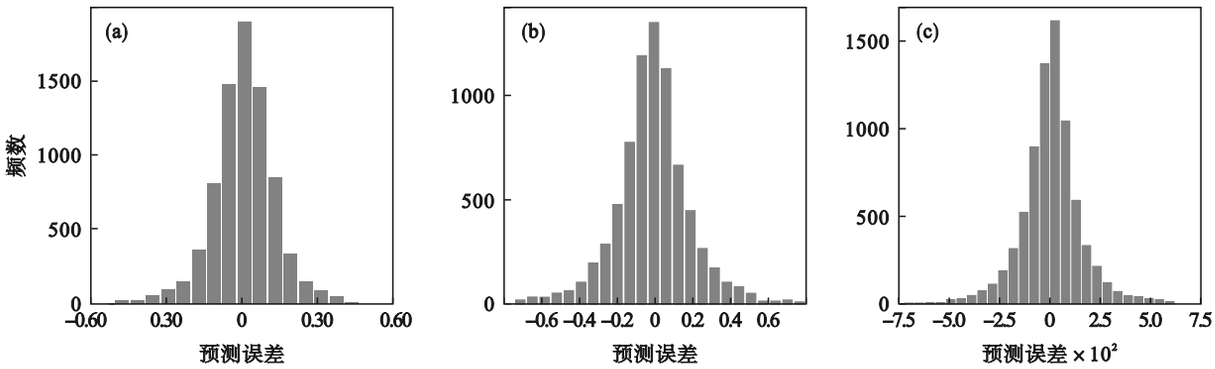

图6　预测误差频数直方图

Fig. 6　Frequency histogram of prediction error

(a)—总铁质量分数;(b)—FeO质量分数;(c)—碱度.

由图6可知,3种化学指标的预测误差都集中在0附近,具有较高的可靠性. 定义预测命中率为预测误差在允许误差范围内的样本数占测试样本总数的比例,BPNN, RNN, LSTM 和GMM−KNN−LSTM模型的预测命中率如图7所示,GMM−KNN−LSTM的预测命中率比其他3种模型的预测命中率高,当总铁质量分数和FeO质量分数预测值在允许误差±0.5%范围内时,GMM−KNN−LSTM预测命中率分别为98.73%和95.33%,比普通的LSTM模型分别提高了3.51%和11%;当碱度预测值在允许误差±0.05范围内时,预测命中率为98.13%,比普通的LSTM模型提高了6.02%,表明了本文所提模型具有较高的预测准确率.

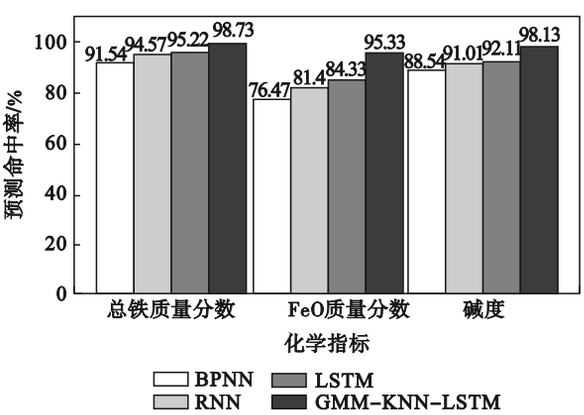

图7　不同模型的预测命中率

Fig. 7　Prediction accuracy of different models

为了更直观地表明GMM−KNN−LSTM模型的预测性能,对比3个化学指标预测值和真实值的分布情况,预测结果如图8所示. 由图8可知,预测值较好地拟合了真实值,本文所建模型具有较好的预测效果.

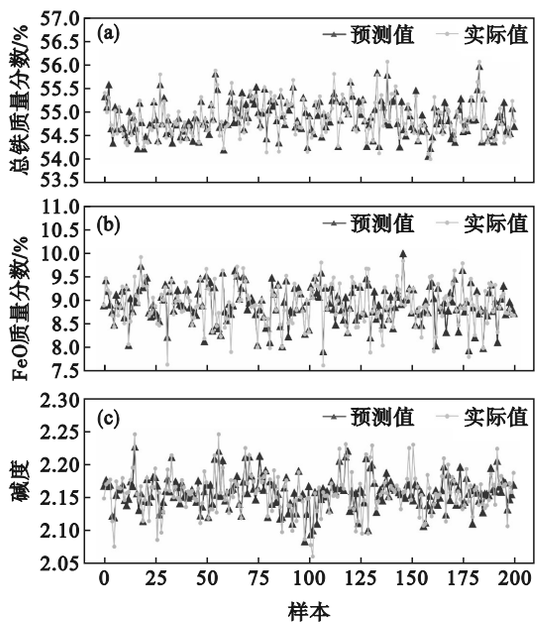

图8　预测值与实际值对比

Fig. 8　Comparison between predicted and actual values

(a)—总铁质量分数;(b)—FeO质量分数;(c)—碱度.

4　结　　论

1)本文首先对烧结过程数据进行了预处理,通过MIC系数法筛选出和3个化学指标强相关的参数,表明了风机风量、风箱负压、风箱废气温度、烧结机速度和原料下料量对化学指标有重要影响,可以作为化学指标预测模型的输入变量. 然后通过GMM−KNN算法使无标签样本参与建模,最后将GMM−KNN算法和LSTM算法相结合,建立了一种基于GMM−KNN−LSTM算法的烧结矿化学指标预测模型.

2)将本文所建模型与未经无标签样本处理的BPNN,RNN和LSTM模型对比,本文所建模型的MSE,MAE和RMSE均比其他3种模型低,

并且误差主要分布在0附近,表明无标签样本中含有的有用信息可以有效提升模型预测精度,解决了无标签样本不能被利用的问题.

3)本文所建模型的总铁质量分数和FeO质量分数预测命中率在允许误差±0.5%范围内时分别为98.73%,95.33%,碱度预测命中率在允许误差±0.05范围内时可达98.13%,表明了本文所建模型具有较高的预测精度,为烧结矿化学指标预测提供了一种可以充分利用烧结过程信息进行建模的有效策略,可以辅助生产人员根据模型的化学指标预测结果来调整配料参数和烧结参数,从而减少烧结矿化学指标的波动.

参考文献:

[1] Liu S, Lyu Q, Liu X J, et al. Synthetically predicting the quality index of sinter using machine learning model[J]. *Ironmaking & Steelmaking*, 2020, 47(7): 828−836.

[2] Li H Y, Li X, Liu X J, et al. Prediction of blast furnace parameters using feature engineering and Stacking algorithm [J]. *Ironmaking & Steelmaking*, 2022, 49(3): 283−296.

[3] 李壮年,储满生,柳政根,等.基于机器学习和遗传算法的高炉参数预测与优化[J].东北大学学报(自然科学版),2020,41(9):1262−1267.
(Li Zhuang-nian, Chu Man-sheng, Liu Zheng-gen, et al. Prediction and optimization of blast furnace parameters based on machine learning and genetic algorithm [J]. *Journal of Northeastern University* (*Natural Science*), 2020, 41(9): 1262−1267.)

[4] Du S, Wu M, Chen L F, et al. Operating mode recognition based on fluctuation interval prediction for iron ore sintering process [J]. *IEEE/ASME Transactions on Mechatronics*, 2020, 25(5): 2297−2308.

[5] Du S, Wu M, Chen X, et al. An intelligent control strategy for iron ore sintering ignition process based on the prediction of ignition temperature[J]. *IEEE Transactions on Industrial Electronics*, 2020, 67(2): 1233−1241.

[6] Jiang Y S, Yang N, Yao Q Q, et al. Real-time moisture control in sintering process using offline-online NARX neural networks[J]. *Neurocomputing*, 2020, 396: 209−215.

[7] Huang Q Y, Liu Z H, Liu Z C, et al. The strength prediction model of iron ore sinter based on an artificial neural network [J]. *Ironmaking & Steelmaking*, 2023, 50(2): 159−166.

[8] Liu S, Lyu Q, Liu X J, et al. A prediction system of burn through point based on gradient boosting decision tree and decision rules[J]. *ISIJ International*, 2019, 59(12): 2156−2164.

[9] Li Z P, Fan X H, Chen G, et al. Optimization of iron ore sintering process based on ELM model and multi-criteria evaluation[J]. *Neural Computing and Applications*, 2017, 28(8): 2247−2253.

[10] 吕庆,刘月明,张振峰,等.基于承钢生产数据预测烧结矿FeO含量[J].钢铁研究学报,2018,30(12):957−962.
(Lyu Qing, Liu Yue-ming, Zhang Zhen-feng, et al. Prediction of FeO content in sinter based on production data of Chengde Steel Mill [J]. *Journal of Iron and Steel Research*, 2018, 30(12): 957−962.)

[11] 刘俊杰,张东升,邵慧君,等.基于BP神经网络的烧结过程预报模型[J].冶金动力,2019,38(1):1−3,9.
(Liu Jun-jie, Zhang Dong-sheng, Shao Hui-jun, et al. Prediction model of sintering process based on BP neural network[J]. *Metallurgical Power*, 2019, 38(1): 1−3, 9.)

[12] Liu S, Liu X J, Lyu Q, et al. Comprehensive system based on a DNN and LSTM for predicting sinter composition[J]. *Applied Soft Computing*, 2020, 95: 106574.

[13] Li D C, Huang W T, Chen C C, et al. Employing box plots to build high-dimensional manufacturing models for new products in TFT-LCD plants [J]. *Neurocomputing*, 2014, 142: 73−85.

[14] Zhou Z, Wen C L, Yang C J. Fault isolation based on *k*-nearest neighbor rule for industrial processes [J]. *IEEE Transactions on Industrial Electronics*, 2016, 63(4): 2578−2586.

[15] 廖志伟,陈琳韬,黄杰栋,等.基于特征空间变换与LSTM的中短期电煤价格预测[J].东北大学学报(自然科学版),2021,42(4):483−493.
(Liao Zhi-wei, Chen Lin-tao, Huang Jie-dong, et al. Medium and short-term electricity coal price forecast based on feature space transformation and LSTM [J]. *Journal of Northeastern University* (*Natural Science*), 2021, 42(4): 483−493.)