DOI: 10.13382/j.jemi.B1902787

面向袋式除尘器的大数据挖掘 XGBoost 优化算法研究*

王梦雅[1] 刘丽冰[1] 熊桂龙[2,3] 赵丹琳[1] 王 宇[1]

(1. 河北工业大学 机械工程学院 天津 300130; 2. 南昌大学 资源环境与化工学院 南昌 330031;
3. 鄱阳湖环境与资源利用教育部重点实验室 南昌 330031)

摘 要:袋式除尘器在产品生命周期不同阶段,包括设计、仿真、制造、测试实验以及运维等阶段都会产生大量数据,挖掘产品大数据与其运行特性之间复杂、非线性、耦合的内在关联,为解决袋式除尘器行业设计创新、运维优化等关键共性技术提供新思路。针对袋式除尘器大数据特点,提出了一种用于袋式除尘器滤袋破损在线监测的大数据挖掘 XGBoost 模型,研究了基于蚁群算法的 XGBoost 模型参数优化方法。研究结果表明,与随机森林、BP 网络挖掘模型相比,XGBoost 优化模型方法准确度高,识别速度快,可解释性强。

关键词: 袋式除尘器;大数据挖掘;XGBoost 模型;蚁群算法优化;破袋监测
中图分类号: TN06;TM925.31 **文献标识码**:A **国家标准学科分类代码**: 460.40

Research on big data mining XGBoost optimization algorithm for bag dust collector

Wang Mengya[1] Liu Libing[1] Xiong Guilong[2,3] Zhao Danlin[1] Wang Yu[1]

(1. School of Mechanical Engineering, Hebei University of Technology, Tianjin 300130, China;
2. School of Resources Environmental and Chemical Engineering, Nanchang University, Nanchang 330031, China;
3. Key Laboratory of Poyang Lake Environment and Resource Utilization, Ministry of Education, Nanchang 330031, China)

Abstract: In different stages of product life cycle, including design, simulation, manufacture, test and operation and maintenance, bag filter generates a large amount of data. It excavates the complex, non-linear and coupling internal relationship between big data of product and its operation characteristics, and provides a new way to solve the common problems of design innovation and operation and maintenance optimization in bag filter industry. Aiming at the characteristics of large data of bag filter, a large data mining XGBoost model for on-line monitoring of bag breakage of bag filter is proposed, and the parameter optimization method of XGBoost model based on ant colony algorithm is studied. Compared with Stochastic Forest and BP network mining models, the results show that the XGBoost optimization model method has high accuracy and strong explanability.

Keywords: bag filter; big data mining; XGBoost model; ant colony algorithm optimization; broken bag monitoring

0 引 言

近年来,随着信息技术的迅速发展和普及应用,各行各业已经进入“大数据时代”,复杂机电产品全生命周期的数据储量与日俱增。挖掘生命周期大数据的真实潜力和价值,使整个生命周期决策更加智能化,是复杂产品制造业发展的未来趋势[1-2]。大数据挖掘作为智能制造最重要的技术之一,可以发现隐藏的知识和其他有用的信息,如生命周期决策和流程参数之间的关系,将大数据挖掘技术应用到生命周期大数据知识发现中,对产品生命周期管理具有重要意义。袋式除尘器作为重要的工业环保装备,除尘效率可达到 99.9%以上,已广泛用于电力、煤炭、钢铁、建材、发电、化工、制药等行业[3-4],积累了大量数据。将这些数据中潜藏的知识挖掘出来,用于指导袋式除尘器整个生命周期的决策,将有利于袋式除尘器

收稿日期:收 2019-11-26 Received Date: 2019-11-26
*基金项目:国家自然科学基金(51666011)、江西省自然科学基金(20171ACB21008)资助项目

产品设计和运行优化。

目前,人工神经网络(ANN)、支持向量机(SVM)、决策树等经典模型已广泛应用于复杂机电产品大数据挖掘中,并取得了一些成果,文献[5]针对空调器长效性能预测中的多因素非线性关系无法确定的问题,建立了基于BP神经网络的空调长效性能预测模型,并通过大数据挖掘的方法将空调在线监测性能与空调结构参数结合起来,确定了空调长效性能达标的最低成本优化策略;文献[6]提出全生命周期设备健康检测诊断的重要性,基于人工神经网络的改进智能算法对设备状态进行了故障预测和定量分析;文献[7]根据电梯当前停层、目的停层以及载荷信息,建立反向传播神经网络(BPNN)模型,来预测电梯行程中吸收或回馈的能量,调节超级电容的平衡电压,从而提前储能或泄能,以补偿电梯运行过程中所需的尖峰功率。文献[8]利用UD-SVR参数寻优对传统的支持向量机模型进行改进,并将改进后的模型用于挖掘输电线路管理系统大数据与输电线路故障诊断中的关联关系;文献[9]针对袋式除尘器数字样机仿真数据,采用分类回归树算法(CART)挖掘设计阶段仿真数据与产品安全性能之间的关系,实现了袋式除尘器安全性能预测。虽然人工智能算法已在大数据挖掘领域有了一定的发展,但多是单模型的机器学习算法的应用,存在预测精度低,容易过拟合,泛化能力较弱的问题。对于袋式除尘器各个生命阶段采集的数据,以上的机器学习算法面对如此大的数据量会存在预测能力不足、运算时间过长的问题。因此,本文提出将集成学习领域中的XGBoost(eXtreme gradient boosting)算法应用于袋式除尘器产品大数据挖掘中。

XGBoost是华盛顿大学的陈天奇博士在2016年基于梯度提升决策树(gradient boosting decision tree, GBDT)算法提出的一种基于梯度提升集成学习算法[10],其原理是通过对多个弱分类器进行集成,经过多次迭代计算得到更加及准确的分类效果。近年来,已有许多将XGBoost算法应用在Kaggle、KDD(knowledge disovery database)等大数据算法竞赛中的实例,由于表现优异,引起了大量关注[11],虽然到目前为止提出只有短短的两年多时间,但其在各个领域已有较多应用[12],Chen等[13]将加权的XGBoost模型用于雷达信号分类;Debaditya等[14]将XGBoost算法应用于建筑能耗预测;也有研究将XGBoost用于电力系统暂态稳态预测[15-16]等。

将XGBoost算法用于袋式除尘器大数据挖掘具有以下优势。

1)XGBoost能够自行采用多线程并行计算,运算速度快,适合处理大规模袋式除尘器全生命周期数据。

2)模型中增加了以树的复杂度构成的正则化项,使得其泛化能力提升,有效解决了过拟合问题。

3)由于XGBoost是树结构模型,不需要对袋式除尘器各个生命阶段采集的各种数据进行归一化处理,并能够有效处理因某些原因导致的缺失值,适用于数据类型多样的袋式除尘器大数据挖掘。

本文以袋式除尘器滤袋破损在线监测为例,研究了用于袋式除尘器大数据挖掘的XGBoost优化模型。首先,针对XGBoost模型多个参数调优效率低、难以找到全局最优解和模型不稳定的问题,采用蚁群算法,利用其强大的全局寻优能力和良好的鲁棒性,对XGBoost模型中的重要参数进行优化;然后,基于优化后的XGBoost模型进行特征重要度排序和筛选;最后,根据所得最优参数和筛选后的重要特征重新训练袋式除尘器滤袋破损状态识别模型,挖掘出袋式除尘器不同生命周期阶段数据与滤袋状态之间的关联关系,实现破袋在线监测。此外,还将蚁群算法与网格搜索算法的参数调优效果进行对比分析;将本文模型与基于随机森林模型、BP神经网络模型的破袋识别效果进行对比分析。

1 袋式除尘器大数据挖掘模型构建

袋式除尘器在其全生命周期的各个阶段积累了大量的数据,其中包括设计过程中的本体结构参数,运维过程中设定的工况参数,以及通过传感器实时高速采集的数据等,这些生命周期数据具有数据量大,实时性高,数据类型多样[17]的特点,符合大数据的"3 V"特性,在对其进行大数据挖掘时。首先,确定挖掘目标,如性能优化或故障诊断;其次,根据挖掘目标,抽取各生命周期阶段中与之相关的参数,将这些参数集成在一起,得到大数据挖掘的样本集;然后,选择合适的数据挖掘方法,建立数据挖掘模型,挖掘出生命周期各阶段数据与袋式除尘器运维特性之间潜藏的关联关系;最后,将挖掘出来的知识应用到袋式除尘器产品生命周期的决策中。这个过程既实现了袋式除尘器生命周期大数据的有效利用,又可将挖掘出的知识用于指导产品设计、运行、维护等阶段的优化,为袋式除尘器行业创新提供了新的思路。

1.1 袋式除尘器大数据挖掘目标的建立

袋式除尘器运行过程中绝大多数处于过滤的状态。因此,袋式除尘器的运行特性的评价指标主要指过滤性能。整机过滤性能主要包括运行能耗、分风均匀性、滤袋破损安全性能3部分。袋式除尘器在运行过程中出现滤袋破损,会导致除尘效率下降,以致设备失效影响生产正常进行,甚至会被环境监管部门强制罚款或予以停产处理,造成重大经济损失。快速准确地检测出除尘器滤袋破损可以及时避免高污染排放,保证生产正常进行。

袋式除尘器破袋检测技术,是一个利用袋式除尘器工作过程中与滤袋破损相关的在线监测数据,通过对数

据的处理,从而判断滤袋是否存在破损的过程。国内外学者对滤袋破损的影响因素问题进行了大量研究,已有研究结果表明,烟气温度、过滤速度、清灰方式等对滤袋破损均有极大的影响[18]。此外,袋式除尘器设备本体参数也对滤袋破损有重要影响,如本体结构参数、进风方式、气流分布、清灰装置、制作和安装质量操作和维护规范等[19],袋式除尘器滤袋破损后,相关传感器数据也会发生改变。

早期,为避免破袋对除尘效率的影响,工厂采用定期更换滤袋的方式来尽量降低破袋发生的可能,但这种方法需要更换大量滤袋,使得除尘器运行成本过高。后来,工厂开始逐步采用人工检漏的方式,但此方法不仅增加了工人的劳动量,而且存在一定的安全隐患,同时影响了袋式除尘器的正常运行。近几年来,国内外学者提出的袋式除尘器滤袋破损检测方法,大多都是通过测量袋式除尘器运行过程中的排放气体的粉尘浓度[20]或测量滤袋内外压差[21]的变化,间接判断滤袋是否破损。这些检测方法检测参数单一,检测方法得出的检测结论比较片面,具有一定的局限性。2019 年,杨宏伟等[22]提出将多传感器信息融合的方法应用到袋式除尘器破袋监测上,给出了基于 D-S 证据理论的滤袋破损监测数据融合方法和实现过程,相比单传感器检测方法,提高了检测的可靠性。但这种方法仍然没有考虑袋式除尘器产品与滤袋相关的结构参数和产品工作时的工况条件,只能监测到破损面积较大的情况,而滤袋具有微小破损时识别能力较低,具有一定的滞后性。

因此,本文以袋式除尘器滤袋破损识别问题作为挖掘目标,将袋式除尘器设计阶段的结构参数、运行阶段的工况参数、传感器参数结合起来,挖掘这些参数与滤袋破损安全性能之间的关联,实现破袋监测。

1.2　袋式除尘器大数据挖掘的 XGBoost 基本模型构建

1)多域数据的获取和处理

袋式除尘器大数据的数据来源范围广,主要通过现场实验和数值模拟获得。其中现场实验依托现场设备,具有工况数据、传感器数据等,数值模拟依托 CFD 及 CAD、CAE、CAT 等 CAX 软件进行数值仿真,包括设计参数(包括 CAD 结构数据和 CAE 工况数据等)、仿真参数、评价数据。本文具体数据来源及数据类型,如表 1 所示。

本文以破袋监测为数据挖掘目标,所以抽取与滤袋性能相关的结构参数(包括滤袋长度、滤袋直径等)、工况参数(过滤速度、工作温度)、传感器数据(滤袋内外压差、花板上下压差、出口粉尘浓度等)、评价数据(滤袋破口大小),将这些数据进行集成,作为模型训练样本集。

2)XGBoost 基本模型构建

基于数值模型采集的袋式除尘器本体结构参数、工况参数、以及传感器数据为模型输入,滤袋破损状态为模型输出。对给定的具有 N 个样本 M 个体征的训练样本集 $D=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}(i=1,2,\cdots,N,\boldsymbol{x}_i\in\boldsymbol{R}^M,\boldsymbol{y}_i\in\boldsymbol{R})$,经过 XGBoost 模型训练,最终得到一个由 K 个 CART 决策树相加的集成模型:

$$\boldsymbol{y}_i^*=\varphi(\boldsymbol{x}_i)=\sum_{k=1}^{K}f_k(\boldsymbol{x}_i),f_k\in F \tag{1}$$

式中:y_i^* 是 XGBoost 模型的预测值输出;$F=\{f(\boldsymbol{x})=w_{q(x)}\}(q:\boldsymbol{R}^M\to T,\boldsymbol{w}\in\boldsymbol{R}^T)$ 为模型中所有 CART 决策树的集合。

XGBoost 模型的损失函数为:

$$L(\varphi)=\sum_{i=1}^{N}l(y_i^*,y_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{2}$$

式中:$l(\,)$ 为训练损失函数,来计算 y_i^* 与 y_i 与之间的偏差(y_i^* 为模型预测值,y_i 为预测值);$\Omega(\,)$ 为正则项。由式(2)可知,损失函数的值应越小越好。添加正则项,可以使模型在保证准确度的同时,不会过于复杂,其定义如式(3)所示。

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda\|\boldsymbol{w}\|^2=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2 \tag{3}$$

式中:T 表示叶子节点数目;$\boldsymbol{w}$ 表示叶子权重;由 γ 和 λ 表示对模型的惩罚系数,一般 λ 取值 1,仅调整参数 γ。

XGBoost 模型训练时,根据式(1),决策树函数 f 逐步增加。假定第 t 步时对第 i 个样本的预测值是 $y_i^{*(t)}$,这时需要增加一个 f_t,来对目标函数进行优化。

$$L^{(t)}=\sum_{i=1}^{N}l(y_i^{*(t)},y_i)+\sum_{K=1}^{t}\Omega(f_k)=$$

$$\sum_{i=1}^{N}l(y_i^{*(t-1)}+f_t(x_i),y_i)+\Omega(f_t)+C \tag{4}$$

其中,C 为第 t 步之前的正则项,是一个常数。此时,新的预测输出变为 $y_i^{*(t-1)}+f_t(x_i)$。将损失函数 $L^{(t)}$ 的减小幅度最大,作为选取树结构 f_t 的标准。将式(4)以二级泰勒级数的形式展开为:

表 1　袋式除尘器具体数据来源及数据类型

Table 1　Specific data source and data type of bag filter

数据来源	数据类型
CAD 结构数据	袋式除尘器设计阶段的结构参数主要包括:滤袋直径及数目、喷吹箱设计参数、分风板设计参数、上箱体设计参数、中箱体设计参数等
CAE 工况数据	袋式除尘器运行阶段的工况参数主要包括:操作参数、介质属性、滤料材质等。其中操作参数有工作温度、过滤速度、出口负压等;介质属性包括介质类型、介质密度、动力粘度等;滤料材质有滤料类型、压力跳跃系数、滤袋厚度、滤袋渗透率等
传感器数据	传感器实时监测设备状态参数,主要有滤袋内外压差、花板上下压差、出口粉尘浓度等
评价数据	评价数据主要用于反映袋式除尘器产品性能的效果,如分风均匀性的评价数据风量相对均方根,可以用来评判风量分配是否均匀;滤袋破损与否、破口面积的大小,用于评判滤袋破损等级

$$L^{(t)} \approx \sum_{i=1}^{N}\left[l(y_i^{*(t-1)},y_i) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) + C \tag{5}$$

式中：$g_i = \partial_{y_i^{*(t-1)}} l(y_i^{*(t-1)},y_i)$，$h_i = \partial^2_{y_i^{*(t-1)}} l(y_i^{*(t-1)},y_i)$ 分别为损失函数 $l(\,)$ 在展开点 $y_i^{*(t-1)}$ 处的一阶导数和二阶导数；由于 $l(y_i^{*(t-1)},y_i)$ 是第 t 步之前的损失函数，是一个定值。故去掉式(5)的常数项，则此时的目标函数 $L_\beta^{(t)}$ 如下：

$$L_\beta^{(t)} \approx \sum_{i=1}^{N}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t) \tag{6}$$

定义 $I = \{i \mid q^t(x_i) = j\}$ 为所有树结构 q^t 映射到第 j 个节点的样本编号集合，则 $L_\beta^{(t)}$ 可进一步被化简。

$$L_\beta^{(t)} \approx \sum_{j=1}^{N}\left[\left(\sum_{i\in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T \tag{7}$$

将该式对 w_j 求导，并且令其为0，即：

$$\sum_{i\in I_j} g_i + \left(\sum_{i\in I_j} h_i + \lambda\right) w_j^* = 0 \tag{8}$$

其最优的叶节点权重为：

$$w_j^* = -\frac{\sum_{i\in i_j} g_i}{\sum_{i\in i_j} h_i + \lambda} \tag{9}$$

将其代入损失函数式(7)中，得到此时最优损失函数：

$$L_{opt\beta}^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{10}$$

$L_{opt\beta}^t$ 用于衡量任意树结构 q^t 的好坏，$L_{opt\beta}^t$ 越小，说明树结构 q^t 使模型的损失函数下降越多，树结构越好。

因此，XGBoost 的训练过程总结如下：以迭代的方式增加 CART 函数，最终获得 XGBoost 模型 $\varphi(x_i) = \sum_{k=1}^{K} f_k(x_i)$。迭代终止的条件为继续增加树模型时，模型准确率提升小于 s。每次增加的新的函数 f_t 获得过程如下：初始有一个叶节点，每次增加一个分叉，选取 $L_{opt\beta}^t$ 最小的树增长方案，循环进行；树停止分裂具有依据下面两个条件：树的最大深度 $d_{\max}$ 达到规定值或最小的样本权重和 $\sum h_i$ 小于设定阈值。树停止分裂后，计算此树结构 q^t 对应的最优权重向量 $\boldsymbol{w}$，从而可得到新的树函数 f_t。

2　蚁群算法对 XGBoost 调参

XGBoost 算法中有许多参数，为了得到良好的分类结果，进行参数寻优是十分必要的。常用的调参方法为网格搜索调优，这种方法是先根据经验固定几个参数，然后对没进行固定参数(一般为 1 个或 2 个)进行网格搜索，再对其他参数依次执行以上操作，最后得到一组优化参数。这种方法一般需要人为监控，得到的参数也不具有全局最优性，且遍历网格内所有参数会相当耗时[23]。由于蚁群算法拥有强大的全局寻优能力，可有效解决参数寻优无法得到全局最优解的问题；而且，蚁群算法的灵活搜索，可有效提高搜索效率；另外，蚁群算法拥有鲁棒性高的特点，在模型构建过程中，存在样本数据含有误差的情况，利用蚁群算法对模型参数进行寻优，可提高模型的稳定性。且采用蚁群算法对进行优化，不会给原本 XGBoost 算法增加很大的复杂性。因此，本文采用蚁群优化算法对 XGBoost 算法进行调参。

2.1　XGBoost 算法参数

XGBoost 参数分为通用参数、Booster 参数以及学习目标参数 3 类，其中 Booster 参数是数据样本进行训练时的主要参数，调整这些参数对模型准确度影响最大。表 2 为 Booster 参数信息表。

表 2　XGBoost 算法的 Booster 参数信息表

Table 2　Booster parameter information table of XGBoost algorithm

参数	取值范围	信息	默认值
learning_rate	[0,1]	学习率，在更新叶子权重的过程中与其相乘，减少每一步权重，来防止过拟合	0.3
max_depth	[0,∞]	最大树深 $d_{\max}$	6
min_child_weight	[0,∞]	最小叶子权重和 $\sum h_i$	1
gamma	[0,∞]	指定节点分裂所需的最小损失函数下降值	0
subsample	(0,1]	控制每棵树随机采样的比例	1
colsample_bythee	(0,1]	控制每颗树随机采样的列数占比	1
colsample_bylevel	[0,1]	控制的每次分裂采样的列数占比	1
⋮	⋮	⋮	⋮

根据大量的 XGBoost 调参经验及工程实践应用，发现学习率(learning_rate)、最大树深(max_depth)、最小叶子样本权重和(min_child_weight)三个参数在模型中的作用明显。过大的 learning_rate 会使算法无法收敛，过小的 learning_rate 又会导致算法过拟合。max_depth 过大，导致模型陷入局部最优解的可能性也变大，从而出现过

拟合现象。min_child_weight 是子节点中最小的样本权重和阀值,该参数过小,会导致算法过拟合,过大则会使算法对线性不可分数据的分类性能降低[24]。因此本文对 learning_rate、max_depth 和 min_child_weight 进行参数寻优;其他参数设置为默认值。

2.2　基于蚁群算法的 XGBoost 参数优化

本文采用蚁群算法建立 XGBoost 算法的参数选择模型。首先,根据优化目标选择合适的目标函数;然后,采用蚁群算法搜寻最优的目标函数值,最后,输出目标函数取得最优函数值所对应的参数取值。优化流程如图 1 所示。其中 x_1 = learning_rate, x_2 = max_depth, x_3 = min_child_weight。

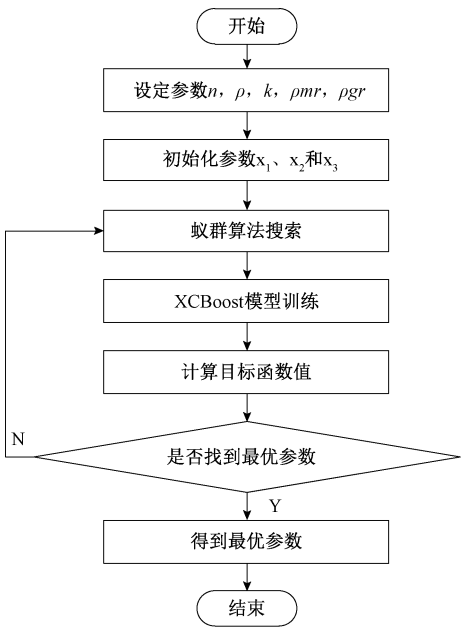

图 1　基于蚁群算法的 XGBoost 模型参数优化流程

Fig. 1　Flow chart optimization of XGBoost model parameters based on ant colony algorithm

1)目标函数的选择

针对 XGBoost 分类问题,目的是得到分类准确率最高的分类模型,因此,以预测集的分类正确率来描述 XGBoost 模型的好坏程度,即:

$$Accuracy = \frac{num(y_i = \varphi(x_i))}{N} \tag{11}$$

式中:y_i 为预测样本所对应的实际值;$\varphi(x_i)$ 为 XGBoost 模型预测值;N 为预测样本个数。由此,得到的的目标函数为:

$$F = \max f(z_1, \cdots, z_j) = \max Accuracy = \max \frac{num(y_i = \varphi(x_i))}{N}$$

$$a_i \leqslant z_i \leqslant b_i; i = 1,2,3 \tag{12}$$

其中,优化变量 z_i 总共为 3 个,对应于参数 learning_rate, max_depth 和 min_child_weight;$[a_i, b_i]$ 为各变量 z_i 的定义域。目标函数即选取最优的参数组合,使测试集的正确率最高。

2)蚁群搜索操作步骤

(1) 蚁群算法参数初始化

初始化蚁群算法的基本参数(种群大小 m、信息素更新比例 ρ 等),设定蚁群搜索的终止条件,本文中设定的终止条件为最大循环次数 $N_{\max}$。

(2) 节点表示及蚂蚁路径的生成

设定 learning_rate, max_depth 和 min_child_weight 3 个参数的有效数位,使 X 代表着 3 个变量的集合,即 $X = \{x_1, x_2, \cdots, x_n\}$,并将各个分量在其定义域等分成 N 个点,共有 $N \times n$ 个点,本文用 $Knot(x_i, y_{i,j})$ 表示一个节点。$x_i (i = 1,2,\cdots,n)$ 定义域内等分的 N 个点所对应的 $Knot(x_i, y_{i,j})$ 节点组成一个平面层 L,共有 n 层。

搜索过程中含有 m 只蚂蚁,每只蚂蚁 k 对应一维数组 $Path_k$(此数组具有 n 个元素),在 $Path_k$ 中存放第 k 只蚂蚁 $(k = 1,2,\cdots,m)$ 依次从 L_1 层到 L_n 层时经过每层对应的纵坐标值,这个过程即为蚂蚁路径的生成。

(3) 迭代搜索过程

①m 只蚂蚁从起点出发,根据式(13)计算每只蚂蚁 $k(k = 1,2,\cdots,m)$ 从 L_{i-1} 层向 L_i 层的转移概率 $P_k(x_i, y_{i,j})$。

$$P_k(x_i, y_{i,j}, t) = \frac{\tau^a(x_i, y_{i,j}, t)\eta^\beta(x_i, y_{i,j}, t)}{\sum_{j=0}^{N} \tau^a(x_i, y_{i,j}, t)\eta^\beta(x_i, y_{i,j}, t)} \tag{13}$$

其中,$\tau(x_i, y_{i,j}, t)$ 为 t 时刻 $Knot(x_i, y_{i,j})$ 上的遗留信息素浓度值,刚开始搜索时,各节点的信息素浓度 $\tau(x_i, y_{i,j}, t) = \gamma(\gamma$ 为常数),信息素浓度增量为 0,即 $\Delta\tau(x_i, y_{i,j}, t) = 0$;$\eta$ 为 $Knot(x_{i-1}, y_{i,j})$ 到 $Knot(x_i, y_{i,j})$ 的期望。

应用赌轮法选取下一步转移的节点(L_i 层某个节点 $Knot(x_i, y_{i,j})$)此时 $Path_k$ 的第 i 个元素即确定为此节点的纵坐标。

②n 个时间单位后,m 只蚂蚁完成一次循环过程中路径爬行,此时,蚂蚁 $k(k = 1,\cdots,m)$ 对应的数组 $Path_k$ 组成一个解 X^*,将其中代表 max_depth 的值向下取整,将这些参数代入 XGBoost 模型中,经过模型训练和模型预测的过程,最终计算出这个解 X^* 对应的模型预测正确率 F^*,并将每只蚂蚁所对应的 m 个模型预测正确率 F^* 进行比较,从而确定本次循环中的最优路径(即最大 F^* 所对应的路径),并记录与之对应的 learning_rate, max_depth 和 min_child_weight 值。

③令 $t = t + n$, $N = N + 1$, 按照式(14)更新各节点上的信息量,并把 $Path_k (k = 1,\cdots,m)$ 中的元素作清零处理。

$$\tau(x_i,y_{i,j},t+n)=\rho\tau(x_i,y_{i,j},t)+\Delta\tau(x_i,y_{i,j})$$

$$\Delta\tau(x_i,y_{i,j})=\sum_{k=1}^{m}\Delta\tau_k(x_i,y_{i,j})$$

$$\Delta\tau(x_i,y_{i,j})=\begin{cases}Q/F_k, & \text{第 } k \text{ 只蚂蚁在本次循环中经过 } Knot(x_i,y_{i,j})\\ 0, & \text{其他}\end{cases} \tag{14}$$

式中：Q 为信息强度；F_k 由式(12)计算。

(4)终止搜索

蚁群算法初始参数设置时，设置的终止条件为蚁群算法最大迭代次数 N_{max}。当迭代次数小于 N_{max} 时，若整个蚁群未收敛到同一路径，则需要将蚂蚁再次置于起始点，重复搜索步骤，直到全部蚂蚁收敛到同一条路径时，算法终止，输出最优路径，即对应的 learning_rate，max_depth 和 min_child_weight.

3 实验验证

本文通过袋式除尘器数值模拟实验模拟了袋式除尘器运行过程中滤袋破损的 4 种状态，滤袋完好、预警状态破损、小面积破损、大面积破损。

1)滤袋完好，破损面积为 0。

2)破袋预警，破损面积属于(0,300] mm^2。

3)小面积破损，破损面积属于(300,1 200] mm^2。

4)大面积破损，破损面积大于 1 200 mm^2。

将实验得到的滤袋参数和与之对应的滤袋破损状态集成，用于后续的大数据挖掘。为了对比本文蚁群算法优化前后的模型的准确性，本文使用 176 个样本数据，将其分为训练集和测试集。

1)训练集包括 105 条数据，训练集中无破损的数据有 11 条，预警状态破损有 37 条，小面积破损有 36 条，较大面积破损有 21 条。

2)测试集包括 71 条数据，测试集中无破损的数据有 9 条，预警状态破损有 23 条，小面积破损有 26 条，较大面积破损有 13 条。

部分实验数据集如表 3 所示。

表 3 部分滤袋相关数据集成

Table 3 Partial filter bag related data integration

滤袋个数	滤袋直径/m	滤袋长度/m	导流板个数	温度/℃	过滤速度/(m·min^{-1})	浓度/(mg·m^{-3})	花板上下压差/Pa	滤袋上内外压差/Pa	滤袋下内外压差/Pa	破损状态
2	0.16	10	0	160	1	0	−900	−1080	−920	0
2	0.16	10	0	160	1.2	0	−930	−1130	−960	0
2	0.16	10	0	160	1	3.9	−588	−568	−410	1
3	0.16	10	0	160	1	3.4	−588	−568	−410	1
3	0.16	10	0	160	1	21.4	−509	−495	−340	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	0.16	10	0	160	1.4	491.3	−289	−274	−90	3

3.1 参数调优

通过蚁群优化算法，对 XGBoost 模型进行参数调优，得到优化后的学习率 learning_rate 为 0.950 462 98，最大树深 max_depth 为 7，最小叶子权重 min_child_weight 为 1.865 160 67；与基于原始的 XGBoost 滤袋破损状态识别模型相比，模型参数优化后，滤袋破损状态识别准确率由起初的 91.55%提高到 97.18%，前提高了 6.16%，如表 4 和 5 所示。参数优化后的 XGBoost 模型具有较好的滤袋破损状态识别性能。

表 4 XGBoost 算法优化前准确率

Table 4 XGBoost algorithm pre-adjustment accuracy

类别	滤袋破损状态	准确率/%	总准确率/%
0	无破损	100	91.55
1	预警状态破损	91.30	
2	小面积破损	88.46	
3	较大面积破损	92.31	

表 5 蚁群算法优化 XGBoost 算法后准确率

Table 5 Ant colony algorithm optimizes XGBoost algorithm accuracy

类别	滤袋破损状态	准确率/%	总准确率/%
0	无破损	100	97.18
1	预警状态破损	100	
2	小面积破损	96.15	
3	较大面积破损	92.31	

3.2 特征筛选

根据本文 3.1 节蚁群算法确定的最优参数组，再次进行 XGBoost 优化模型训练。XGBoost 模型集成若干棵回归树后，每棵树的节点都是在做一次特征分裂，可以将某特征被选为分裂特征的次数作为该特征的重要度。即若一个特征作为分裂特征的次数越多，则这个特征对于滤袋破损状态分类越重要，因此，可以得到所有特征的重要度排序。本文中滤袋破损状态识别中输入特征的重要度如图 2 所示。

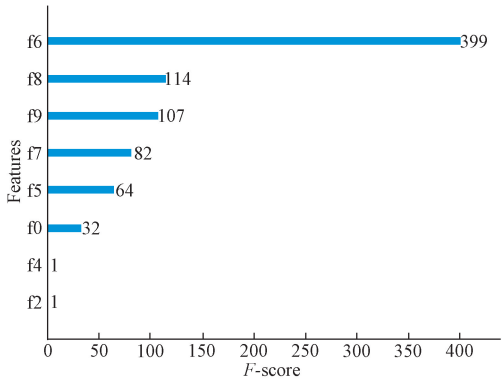

图 2　滤袋破损状态识别中输入特征的重要度排序

Fig. 2　Order of importance of input features in filter bag damage state recognition

图 2 纵坐标的 f0~f9 分别为滤袋个数、滤袋直径、滤袋长度、导流板个数、温度、过滤速度、出口粉尘浓度、花板上下压差、滤袋上内外压差、滤袋下内外压差。图 2 横坐标的 *F*-score 越大，意味着对应的特征越重要；从图 2 可知，前 6 个重要度较大的元素分别为出口粉尘浓度、滤袋上内外压差、滤袋下内外压差、花板上下压差、过滤速度、滤袋个数。

因此，本文选用这 6 个元素作为模型输入对滤料破损状态进行识别。

3.3　XGBoost 模型训练及结果分析

根据 3.1 与 3.2 节得到的 XGBoost 最优参数组和重要特征，建立袋式除尘器滤袋破损在线监测的 XGBoost 模型，将测试数据输入模型进行识别，判断滤袋破损状态，并计算模型对测试数据集识别的准确度。表 6 为这 6 个特征的表现效果，整体预测的准确率不但没有降低，而且有效提高了模型预测效率。

表 6　特征筛选前后模型对比

Table 6　Model comparison before and after feature screening

类别	10 个输入参数	6 个输入参数
准确率/%	97. 18	97. 18
运行时间/s	0. 107	0. 048

4　实验对比

4.1　蚁群算法与网格搜索算法进行参数调优的对比

在参数调优对比实验中网格搜索算法将 XGBoost 中的 3 个参数值(学习率、最大树深和最小叶子权重)的可行区间按从大到小的顺序排列并划分出一些小区间，由计算机按顺序计算各参数变量值组合所对应的正确率，从而求得本区间最大目标值和其对应的最佳参数值。本文设计了单参数网格搜索形式与全参数网格搜索形式，其中单参数网格搜索形式只改变目标参数值，其他参数采用模型的默认值；而全参数网格搜索形式对所有参数进行排列组合，缺点是模型搜索的时间较长。

表 7 为单参数网格搜索算法、全参数网格搜索算法与本文采用的蚁群优化算法对这 3 个参数的调优结果及调参后准确率对比，可以看出全参数网格搜索算法 3 个参数值除最小叶子权重不同，其余两个参数值已经非常接近蚁群算法优化出的参数值，但最后优化模型的准确率还是本文采用的蚁群算法参数调优最高，达到 97. 18%。

表 7　本文调优模型与网格搜索算法参数调优对比

Table 7　Comparison between the optimization model and grid search algorithm

	本文模型	单参数网格搜索	全参数网格搜索
学习率	0. 950 462 98	0. 111 1	1
最大树深	7	6	7
最小叶子权重	1. 865 160 67	3	5
准确率/%	97. 18	92. 96	94. 37

4.2　XGBoost 优化模型与其他分类模型对比

为了测试 XGBoost 优化模型的准确性和运行效率，将其与另一种组合分类算法随机森林，以及经典单模型中应用较广泛的算法 BP 神经网络进行比较。

表 8 为基于优化的 XGBoost 模型、随机森林、BP 神经网络模型的理论研究结果，对比分析可知，本文优化的 XGBoost 模型无论是在准确率还是运行时间方面，均优于另外两种模型。

表 8　与其他两种模型的识别准确率对比

Table 8　Comparison with the recognition accuracy of the other two models

	本文模型	随机森林	BP 神经网络
0 的准确率/%	100	100	100
1 的准确率/%	100	82. 61	82. 61
2 的准确率/%	96. 15	96. 15	100
3 的准确率/%	92. 31	84. 62	92. 31
总准确率/%	97. 18	90. 14	92. 96
运行时间/s	0. 048	3. 5450	0. 910

5　结　论

本文结合袋式除尘器大数据的特点，提出了一种基于 XGBoost 优化模型的袋式除尘器破袋监测方法，挖掘袋式除尘器生命周期大数据与其滤袋破损安全性能的内在关联，该方法采用蚁群算法，对 XGBoost 模型中的重要参数进行优化，改善了传统 XGBoost 参数寻优过程中存在的多参数调优速度慢、易陷入局部最优解的问题，提高了预测精度；根据特征重要度排序挖掘出特征与滤袋破损状态之间的关系，利用特征重要度排序筛选重要特征，

加快了模型训练速度。相对于传统监测方法,该方法能够综合考虑产品各生命周期相关参数对滤袋破损的影响,实现破袋状态的准确监测,同时使监测结果具有了实时性,甚至可预测性,有望应用于实际除尘器产品中。

参考文献

[1] REN S, ZHANG Y, LIU Y, et al. A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: A framework, challenges and future research directions[J]. Journal of Cleaner Production, 2019, 210: 1343-1365.

[2] OPRESNIK D, TAISCH M. The value of Big Data in servitization [J]. International Journal of Production Economics, 2015, 165:174-184.

[3] 吴佳林, 郝俊强, 凡祖伟. 袋式除尘守护蓝天[J]. 纺织科学研究, 2014(8):18-21.
WU J L, HAO J Q, FAN Z W. Bag Dust Removal and Protecting the Blue Sky[J]. Textile Science Research, 2014(8): 18-21.

[4] 熊桂龙, 李水清, 陈晟, 等. 增强 PM2.5 脱除的新型电除尘技术的发展 [J]. 中国电机工程学报, 2015, 35(9):2217-2223.
XIONG G L, LI SH Q, CHEN SH, et al. Development of advanced electrostatic precipitation technologies for reducing $PM_{2.5}$ emissions from coal-fired power plants [J]. Proceedings of the CSEE, 2015, 35(9):2217-2223.

[5] 巫江虹, 刘超鹏, 梁志豪, 等. 房间空调器长效运行性能预测及优化方案的研究[J]. 机械工程学报, 2015, 51(18):158-166.
WU J H, LIU CH P, LIANG ZH H, et al. Research on long-term operation performance prediction and optimization scheme of room air conditioner[J]. Journal of Mechanical Engineering, 2015, 51(18): 158-166.

[6] 吴天舒, 陈蜀宇, 吴朋. 全生命周期健康监测诊断系统研究[J]. 仪器仪表学报, 2018,39(8):204-211.
WU T SH, CHEN SH Y, WU P. Research on the whole life cycle health monitoring and diagnosis system [J]. Chinese Journal of Scientific Instrument, 2018, 39(8): 204-211.

[7] 张达敏, 林辉品, 林智勇, 等. 基于神经网络预测控制的节能电梯能量管理[J]. 仪器仪表学报, 2017, 38(12):3137-3142.
ZHANG D M, LIN H P, LIN ZH Y, et al. Energy-saving elevator energy management based on neural network predictive control [J]. Chinese Journal of Scientific Instrument, 2017, 38(12): 3137-3142.

[8] 李志鹏. 基于大数据分析的输电线路管理系统及故障诊断研究[D]. 武汉:湖北工业大学, 2015.
LI ZH P. Transmission line management system and fault diagnosis based on big data analysis[D]. Wuhan: Hubei University of Technology, 2015.

[9] 康学娟. 袋式除尘器数字样机数据挖掘技术及应用研究[D]. 天津:河北工业大学. 2015.
KANG X J. Data mining technology and application research of digital dust collector for bag filter [D]. Tianjin: Hebei University of Technology, 2015.

[10] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system [C]. Proceedings of the 22nd ACM SICKDD International Conference on Knowledge Discovery and Data Mining, 2016:785-794

[11] DIDRIK N. Tree Boosting with XGBoost - why does XGBoost win "every" machine learning competition? [D]. Norway: Norwegian University of Science and Technology, 2016.

[12] 伯毅. 基于 XGBoost 模型的短期股票预测[D]. 哈尔滨:哈尔滨工业大学. 2018
BO Y. Short-term stock forecast based on XGBoost model [D]. Harbin: Harbin Institute of Technology. 2018

[13] CHEN W, FU K, ZUO J, et al. Radar emitter classification for large data set based on weighted-XGBoost [J]. Iet Radar Sonar & Navigation, 2017, 11(8): 1203-1207.

[14] DEBADITYA C, HAZEM E. Advanced machine learning techniques for building performance simulation: a comparative analysis [J]. Journal of Building Performance Simulation, 2019, 12(2): 193-207.

[15] 陈明华,刘群英,张家枢,等. 基于 XGBoost 的电力系统暂态稳定预测方法研究[J]. 电网技术, 2020, 44(3): 1026-1034.
CHEN M H, LIU Q Y, ZHANG J SH, et al. Research on power system transient stability prediction method based on XGBoost [J]. Grid Technology, 2020, 44(3): 1026-1034.

[16] 张晨宇, 王慧芳, 叶晓君. 基于 XGBoost 算法的电力系统暂态稳定评估[J]. 电力自动化设备, 2019, 39(3):83-89,95.
ZHANG CH Y, WANG H F, YE X J. Power system transient stability assessment based on XGBoost algorithm[J]. Power Automation Equipment, 2019, 39(3): 83-89,95.

[17] 任杉, 张映锋, 黄彬彬. 生命周期大数据驱动的复杂产品智能制造服务新模式研究[J]. 机械工程学报, 2018, 54(22):208-217.
REN SH, ZHANG Y F, HUANG B B. New model of intelligent manufacturing service for complex products driven by large data in life cycle [J]. Journal of Mechanical Engineering, 2018, 54 (22): 208-217.

[18] 王丹丹, 钱付平, 夏勇军, 等. 基于故障树分析法袋

式除尘器滤袋失效的研究与应用[J]. 环境工程学报, 2016, 10(6):3118-3124.
WANG D D, QIAN F P, XIA Y J, et al. Research and application of bag filter failure based on fault tree analysis method [J]. Journal of Environmental Engineering, 2016, 10(6): 3118-3124.

[19] 陈隆枢. 设备本体缺陷导致袋式除尘器失效因素探析[J]. 中国环保产业, 2012(3):34-39.
CHEN L SH. Analysis of failure factors of bag dust collector caused by defects in equipment body[J]. China Environmental Protection Industry, 2012(3): 34-39.

[20] 刘艳. 脉喷袋除尘器智能破袋监测可行性分析[J]. 中国水泥, 2015(11):99-100.
LIU Y. Feasibility analysis of intelligent bag breaking monitoring of pulse jet bag filter [J]. China Cement, 2015 (11): 99-100.

[21] 余新明, 吴学军, 吕先昌. 布袋收尘穿漏监测及定位技术现状与展望[J]. 工业安全与环保, 2005, 31(5):13-14.
YU X M, WU X J, LV X CH. Current situation and Prospect of bag dust collection and leakage monitoring and positioning technology [J]. Industrial Safety and Environmental Protection, 2005, 31 (5): 13-14.

[22] 杨宏伟,熊桂龙,张松,等. 多传感器信息融合滤袋破损检测方法[J]. 河北工业大学学报,2019,48(6):6-11.
YANG H W, XIONG G L, ZHANG S, et al. Detection method of multi-sensor information fusion filter bag damage [J]. Journal of Hebei University of Technology, 2019,48 (6): 6-11.

[23] 刘佳, 施龙青, 韩进, 等. 基于 Grid-Search_PSO 优化 SVM 回归预测矿井涌水量[J]. 煤炭技术, 2015, 34(8):184-186.
LIU J, SHI L Q, HAN J, et al. Prediction of mine water inflow based on Grid-Search _ PSO optimized SVM regression [J]. Coal Technology, 2015, 34 (8): 184-186.

[24] DIAZ-MORALES R, NAVIA-VAZQUEZ A. Optimization of AMS using weighted AUC optimized models [J]. JMLR W&CP, 2015, 42: 109-127.

作者简介

王梦雅,2017 年于河北工业大学获得学士学位,现为河北工业大学硕士研究生,主要研究方向为工业测控技术。
E-mail:947745509@ qq. com

Wang Mengya received her B. Sc. degree from Hebei University of Technology in 2017. Now She is a M. Sc candidate at Hebei University of Technology. Her main research direction is industrial measurement and control technology.

刘丽冰,1983 年于河北工业大学获得学士学位,1995 年于河北工业大学获得硕士学位,1998 年于天津大学获得博士学位,现为河北工业大学博士生导师,主要研究方向为数字化集成制造技术及装备、智能测控技术、复杂系统模型及应用。
E-mail:tjxiaobing@ 163. com

Liu Libing received her B. Sc. degree from Hebei University of Technology in 1983, M. Sc. degree from Hebei University of Technology in 1995, and Ph. D. from Tianjin University in 1998. Now she is a Ph. D. supervisor at Hebei University of Technology. Her main research interests include digital integrated manufacturing technology and equipment, intelligent measurement and control technology, complex system model and application.

熊桂龙(通信作者),2012 年于东南大学获得博士学位,2012~2015 年于清华大学从事博士后研究,现为南昌大学助理研究员,硕士生导师,主要研究方向为燃煤污染物排放控制,除尘技术、多相流动与数值模拟研究。
E-mail:jcijxcn@ 163. com

Xiong Guilong(Corresponding author) received his Ph. D. degree from Southeast University in 2012, and worked as a postdoctoral student at Tsinghua University in 2012-2015. Now he is an assistant researcher and M. Sc. supervisor at Nanchang University. His main research interests include emission control of coal-fired pollutants, dust removal technology, multiphase flow and numerical simulation.