

DOI: 10.13382/j.jemi.B2205797

基于主动学习和最优路径森林的 管道故障分类识别方法^{*}

李 恬^{1,2} 冯 早^{1,2} 朱雪峰^{1,2}

(1. 昆明理工大学信息与自动化学院 昆明 650500; 2. 云南省人工智能重点实验室 昆明 650500)

摘 要: 在工业故障分类过程中有标记样本数量少而人工标注成本高会导致分类器精度难以提高,而大量包含丰富信息的无标记样本却没有得到充分利用。针对上述问题,提出了一种结合主动学习(AL)和最优路径森林算法(OPF)的半监督故障分类模型(AL-OPF)。该方法首先利用BvSB和余弦相似度准则综合衡量样本的价值量,以排序批处理模式筛选价值高的样本,并获取其标签扩充初始标记样本集,然后通过构建最优路径森林实现半监督标签传播,最后在实验室采集得到的管道故障样本集上进行实验验证。实验结果表明,该方法能在有标签样本为10%的情况下达到96.68%的整体识别准确率,与逐个采样模式的主动学习方法以及基于距离度量提取训练样本全局结构信息的半监督方法相比,所提出方法拥有更高的Recall值和F1-score值。

关键词: 主动学习;最优路径森林;半监督;故障分类
中图分类号: TP274.2 **文献标识码:** A **国家标准学科分类代码:** 510.40

Pipeline blockage recognition method based on active learning and optimum-path forest

Li Tian^{1,2} Feng Zao^{1,2} Zhu Xuefeng^{1,2}

(1. Faculty of Information Engineering & Automation, Kunming University of Science and Technology, Kunming 650500, China;
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming 650500, China)

Abstract: Aiming at the problem of difficulty in improving the classification accuracy of industrial fault detection caused by its limited number of labeled training samples which would consume a significant amount of manpower, which a large number of unlabeled samples containing rich information are not fully utilized, this paper puts forward a semi-supervised classification model of combining active learning (AL) and the optimum-path forest (OPF). Firstly, the high-value samples are selected in sorting batch mode according to the value of samples that are comprehensively measured based on BvSB and cosine similarity criterion, and the value of each sample is obtained to expand the initial labeled sample set. Secondly, semi-supervised label propagation is achieved by constructing the optimum-path forest. Finally, the experimental verification was carried out using laboratory collected pipe condition datasets. The experimental results show that the method can achieve an overall recognition accuracy of 96.68% when the number of labeled samples is 10%. Compared with active learning methods in one-by-one sampling mode and semi-supervised methods that extract global structural information of training samples based on distance metrics, the proposed method has higher Recall value and F1-score value.

Keywords: active learning; optimum-path forest; semi-supervised; fault classification

0 引 言

传统有监督分类算法的性能对有标注样本的数量和质量有较高的依赖性,但是在实际的应用场景中,大量的

有标记数据较难获得,而对样本数据进行标注则需要花费大量的人力和计算资源,针对这类问题,主动学习和半监督学习原理和方法逐渐被应用到故障识别领域。主动学习和半监督学习都是利用少量有标注样本数据和大量无标注数据提升分类模型性能的方法,半监督学习侧重

探索无标注数据中的模型已知部分,而主动学习则尝试挖掘未知的富有信息量的数据,它们具有内在相似性和良好互补性^[1]。

主动学习方法只挑选少量有价值的样本更新模型而忽略了剩余大量无标注样本中包含的信息,而半监督学习可以提取无标注样本的全局结构信息,有助于建立更加精准的分类模型,实现这两类方法的有机组合使用,可以在提高分类模型性能的同时明显减少人工标记的工作量。基于上述思想,近期产出的相关研究成果包括:Jian 等^[2]提出了一种基于主动学习和半监督学习的小训练集工业故障诊断方法,利用主动学习扩充初始训练集后基于协同训练架构提高小训练集下的故障诊断性能;韩彦岭等^[3]针对海冰遥感图像分类问题中标签样本获取困难导致海冰分类精度难以提高的问题,提出了一种融合主动学习选出具有代表性的样本建立半监督分类模型的方法,提高了半监督学习的性能;Han 等^[4]提出了一种在混合环境下基于主动半监督学习的声音分类方法,该方法与其他主动学习相比只需要很少的标注样本就能达到相同的性能;Yin 等^[5]提出了一种基于主动学习的半监督 FDA 工业故障分类模型,使用主动学习标记样本提高模型性能,同时半监督学习利用额外未标记数据减少人工标注成本。

排水管道作为城市的重要组成部分,保障了城市居民生活的正常运转,但随着管道服役周期不断延长,管道会出现老化、腐蚀等状况^[6],导致管道内部产生不同程度的堵塞,如果堵塞不能被及时发现并排除,会造成管道承载能力下降,传输效率降低的同时可能带来安全隐患。由于管道安全运行的需要,国内外学者展开了大量关于管道堵塞检测方法的研究工作,目前主要的检测方法有瞬态分析、声反射、超声波检测等^[7],其中声学主动检测方法作为一种无损的检测方法在故障检测领域中有着检测效率高、不易受检测对象运行状态影响以及检测结果客观性较强等优势。基于超声检测原理,邓勇等^[8]提出一种深度学习的天然气管体压力超声检测模式识别方法实现对管道压力有无的判别;基于声学主动检测原理,焦敬品等^[9]利用 BP 神经网络进行管道泄漏声信号识别方法研究;赵燕锋等^[10]提出了一种基于时频图像和卷积神经网络的管道堵塞物和三通件识别方法,完成对不同堵塞物和三通件的自动识别。以上研究是在有标注故障样本集充足的情况下对运行工况进行识别,而实际管道运行中对故障数据进行标注的代价高昂,因此,如何在有标注样本数量少的限制下充分挖掘大量无标注数据中的潜在信息,实现对管道运行状态的评估是目前研究需解决的问题。

针对管道在实际运行过程中有标记样本获取成本较高而存在大量无标记样本所包含的信息却没有被充分挖

掘的问题,本文提出了一种结合主动学习和最优路径森林算法的半监督故障分类模型。首先利用 BvSB 和余弦相似度准则综合衡量样本的信息量,以排序批处理模式选择具有代表性的无标记样本主动标记后扩充有标记样本集,然后基于样本间的距离与分布构建最优路径森林实现半监督标签传播,在实验室采集得到的管道声学样本集上验证该方法,结果表明利用主动学习标记的样本能够有效改善分类器错误传播标签的问题,在有标注样本集较少的情况下,该方法具有较好的鲁棒性且具有训练时间短,故障分类准确率高等优点。

1 相关理论

1.1 基于 BvSB 和余弦相似度的排序批处理主动学习

主动学习主要包括学习模块和选择模块^[11]。主动学习的关键在于如何从无标注样本集中选择出最有价值的实例并获取它们的标记信息,样本选择准则主要包括:不确定性准则、版本空间缩减准则、泛化误差缩减准则等^[12]。

常规的主动学习方法每次只采样一个样本对模型进行更新,这种方法在样本量较少时能获得较好的结果,但在样本量较大时需要进行多次迭代才能获得较高精度,因此影响算法执行的速度和效率,为解决这个问题,本文采用一种排序批处理^[13]的主动学习采样策略一次采集多个样本,同时,为了避免在挑选样本时选择冗余样本,综合考虑样本的差异性和代表性来提升所筛选样本的质量,在减少算法本身耗时的同时选取价值高的样本,算法流程图如图 1 所示。

设无标注样本集为 $U = \{X_1, X_2, \dots, X_n\}$, $Y = \{1, 2, \dots\}$ 为所有数据包含的全部类别,通过当前有标注样本集训练得到的分类模型,可得到每个样本属于各个类别的概率为 $P(y_i | x_i)$,样本信息度估计采用最优标号与次优标号原则,该方法考虑了待分类样本的类属概率值为最高概率和次高概率之间的差异,数学公式表示如下:

$$x_M = \arg \min_x P(y_{best} | x_i) - P(y_{sec_{ond}} | x_i) \quad (1)$$

为防止所筛选的样本集冗余,引入余弦相似度来衡量所筛选未标注样本的代表性,两个 n 维样本点的夹角余弦公式为:

$$\cos(\theta) = \frac{\sum_{k=1}^n x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^n x_{1k}^2} \cdot \sqrt{\sum_{k=1}^n x_{2k}^2}} \quad (2)$$

通过计算余弦相似度来衡量未标注样本和标注样本的相似性,该公式表示为:

$$d_{\cos}(x_i, x_L) = \frac{1}{k} \sum_{x' \in x_L} \cos(x_i, x') \quad (3)$$

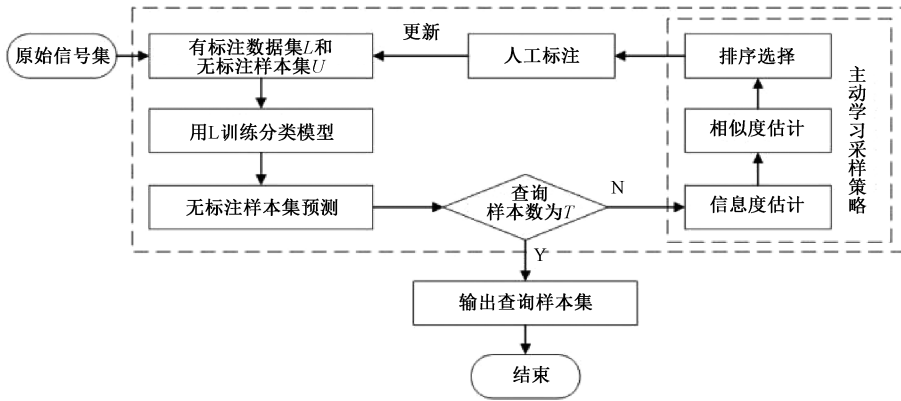


图 1 主动学习算法流程图

Fig. 1 Flow chart of active learning

式中: k 表示已标注样本集样本个数, x_i 表示所选择的未标注样本。综合考虑样本的不确定性和代表性, 采用最优和次优标号原则和余弦相似度相结合的方式查询样本, Q 越大代表样本的价值越高, 计算公式如下:

$$Q = \lambda \cdot x_M + (1 - \lambda) d_{\cos}(x_i, x_L) \quad (4)$$

式中: 参数 λ 设置为 0.5, 表示样本的不确定性和代表性一样重要, 这样所筛选的样本不仅能减少数据集的冗余, 在加入训练集后也能使模型快速收敛。

1.2 基于最优路径森林算法的半监督学习

最优路径森林算法是一种基于完全图的监督分类算法, 该方法包括训练分类器和分类两个阶段^[14], 首先根据所有训练样本之间的距离构建完全图, 基于完全图构成最小生成树, 再根据训练集样本的类别生成对应最优路径树, 然后构建最优路径森林分类器并对测试样本进行分类, 在进行分类时, 根据距离和分布计算待分类样本到哪颗树的距离最近, 则将其归属到这棵树的根节点所属类别^[15]。由于该算法可以包含有未标记的非原型样本, 因此可以扩展到半监督邻域, 与其他基于距离和分布的半监督方法相比, OPF 算法不依赖于任何参数, 训练阶段不需要对参数进行寻优, 能处理多分类及有一定程度类别间重叠的问题, 并且具有精度高、响应速度快以及初始标记训练样本需求不大等优点^[16]。

将总数据集 X 分成训练集 X_1 和测试集 X_2 , 假设 X_1 共有 N 个样本, 每个样本的特征维度为 M , 首先利用 X_1 的 N 个样本构建完全图 A , 图中的每个节点代表集合 X_1 中的一个样本, 图 A 中任意两个节点 j 和 k 之间都有弧, 弧的权值为两个节点之间的欧氏距离, 公式如下:

$$d(x_j, x_k) = \sqrt{(x_{j1} - x_{k1})^2 + \cdots + (x_{jM} - x_{kM})^2} \quad (5)$$

对完全图 A 生成其最小生成树, 在得到的最小生成树中, 连接不同类别节点的弧所对应的节点为最优路径森林中树的根节点, 如图 2(a) ~ (b) 所示。定义路径为单个或多个不同节点构成的连续节点序列 $\pi = (s_1,$

$s_2, \dots, s_k)$, 其中 $s_i \in A$ 且 $1 \leq i \leq k-1$ 。路径的起点为根节点, 终点是任意节点, 根据路径代价函数可以计算每条路径的代价 $f(\pi)$, 若路径是最优的, 则与该路径具有相同终点 s_k 的任何其他路径 π' , 都有 $f(\pi) \leq f(\pi')$ 。根据最优路径森林算法的路径代价函数可求得各节点的代价初始值及路径代价为:

$$f_{\max}(\langle s \rangle) = \begin{cases} 0, & \text{当 } s \text{ 是根节点时} \\ +\infty, & \text{当 } s \text{ 为其他节点时} \end{cases} \quad (6)$$

$$f_{\max}(\pi \cdot \langle s, t \rangle) = \max\{f_{\max}(\pi), d(s, t)\} \quad (7)$$

式中: $\pi \cdot \langle s, t \rangle$ 表示以 s 为终点的路径 π 与弧 $\langle s, t \rangle$ 所构成的路径, $f_{\max}(\pi)$ 的值为路径 π 上任意相邻两点之间的最大距离值。

节点到终点的不同路径中代价最小的路径就是该节点的最优路径, 最优路径森林算法将找到根节点集合 S 到训练集 X_1 中每一个节点 t 的最优路径, 从而构建最优路径森林^[17], 并得到每个节点的最优路径代价值 $C(t)$ 以及该节点所属类别, 节点 t 的代价 $C(t)$ 为最优路径上 t 的前驱节点 s 的代价 $C(s)$ 与节点 s 和 t 之间距离 $d(s, t)$ 的最大值, t 的类别就是其前驱节点所属类别 $L(R(s))$ (其中 $R(s)$ 表示 s 所在最优路径树上的根节点), 如图 2(c) ~ (d) 所示。

对于测试集 X_2 中的待分类样本 $t \in X_2$, 计算 t 到最小生成树中所有节点的距离, 考虑 t 通过 X_1 中所有节点的路径代价, 选择全部路径中 t 的代价最小的路径就是最优路径 $P^*(t)$, t 的类别与该最优路径的根节点的类别相同, 如图 3(e) ~ (f) 所示。

$$L(t) = \lambda(R(t)) \quad (8)$$

$$C(t) = \min\{\max\{C(s), d(s, t)\}\}, \forall s \in X_1 \quad (9)$$

2 基于主动学习的 OPF 组合算法流程

为了尽量减少分类器训练过程中所需人工标记样本

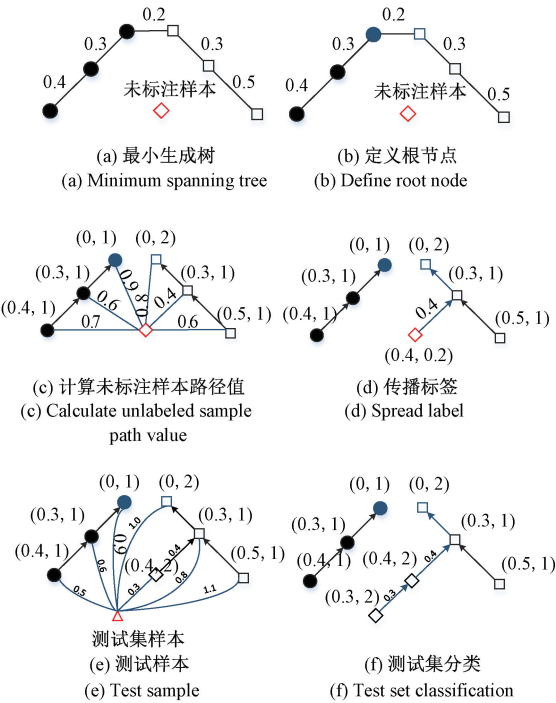


图 2 半监督最优路径森林算法步骤图

Fig. 2 Step diagram of semi-supervised optimum-path forest

的数量,同时获得性能较好的分类器,本文提出利用主动学习和最优路径森林算法相结合构建管道堵塞故障分类识别模型,其执行步骤主要包括数据预处理、BvSB 和余弦相似度的批处理主动学习、OPF 半监督学习以及模型测试,具体算法流程图如图 3 所示。

将构建的声学特征集分为有标注样本集和无标注样本集分别记为 L 和 U ,由用户对有标注样本集进行人工标注,用训练集 L 训练随机森林分类器,同时预测无标注样本集 U 中的样本的类别,并计算其属于各个类别的概率 $P(y_i|x)$, $y_i \in Y, X \in U$ 。利用 BvSB 和余弦相似度主动学习来查询无标注样本集中最有价值的样本,根据上述得到的无标注样本属于各个类别的概率 $P(y_i|x)$,计算 Q 值并排序,从 U 中选出 T 个样本由用户进行人工标注,扩充初始有标注样本集。

将更新后的样本集分为训练集和测试集,利用训练集构建完全图并建立对应的最小生成树,计算每个节点的最优路径以及代价值,构成最优路径森林,直到训练集中无标注样本集为空时停止,最后利用所构建的最优路径森林对测试集进行分类识别。

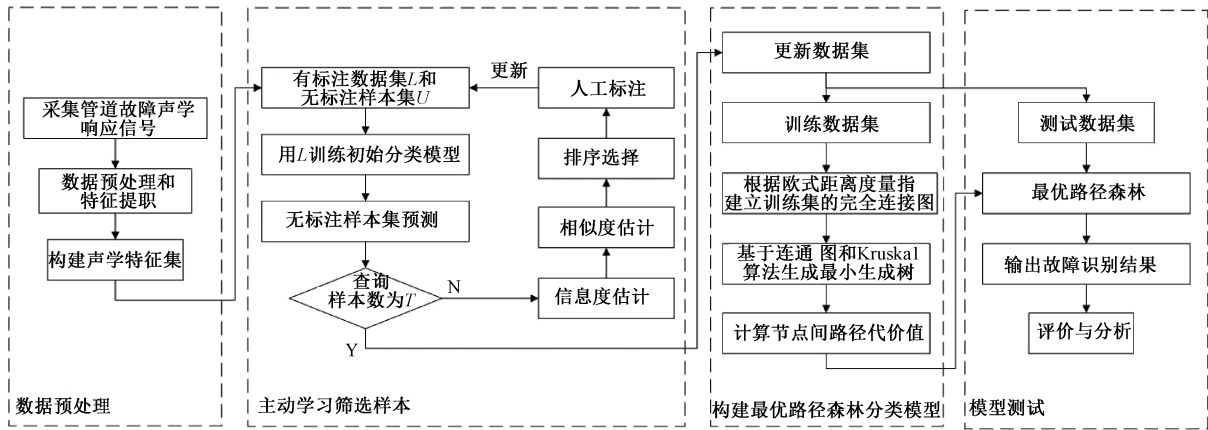


图 3 算法流程图

Fig. 3 Flow chart of algorithm

3 实验验证

以管道堵塞故障为主要研究对象,建立基于声学主动检测原理的管道状态分类识别模型主要分为数据采集和预处理、特征提取和分类识别。

3.1 实验平台搭建和数据采集

为模拟排水管道运行工况,实验室采用长 15.4 m、直径 150 mm 的黏土管道搭建了图 4(a) 所示的实验平台,图 4(b) 所示为实验中的三通件和堵塞模拟物。实验装

置包括计算机、功率放大器、水听器、扬声器、滤波器以及两根伸缩杆,具体数据采集过程如下:计算机控制 WinMLS 软件产生长度为 10 s,频率为 100~6 000 Hz 的正弦扫频信号,信号经由放大器放大后通过扬声器发出声波信号到管道内部,4 通道的水听器采集反射声波通过滤波器滤波后传输到计算机中进行分析处理,实验采样频率设置为 44 100 Hz,管道检测实验平台如图 4(c) 所示。声波在管道中传播时遇到堵塞物后会导致声波扰动,从而引起介质压强变化,压强的变化量描述了声波在管道中的传播特性^[18],声音在管道内的这种特性为声学

主动检测管道堵塞故障提供了依据,通过对管道的声压信号进行处理分析、特征提取和模式识别,进而得到管道运行状况实现对故障的分类。

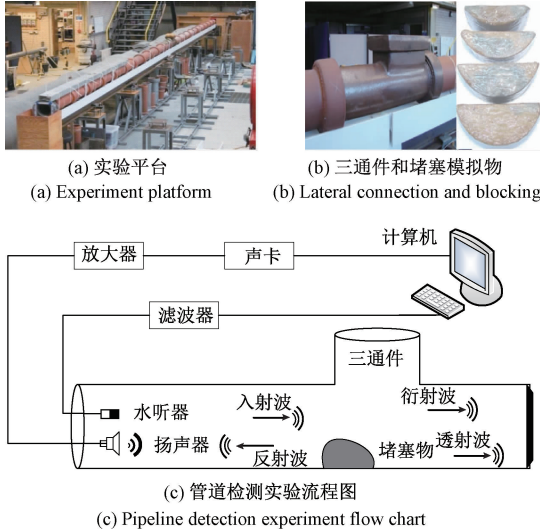


图 4 管道检测实验图

Fig. 4 Pipeline detection experiment chart

实验中堵塞程度的判定是依据堵塞物的高度相对于管道的直径,其中实验室设定管道内部堵塞物的高度低于管道直径 $\frac{1}{3}$ 为轻度堵塞,超过管道直径 $\frac{1}{3}$ 为中重度堵塞,实验设计并模拟了 4 种管道运行工况,包括:1) 正常管道无堵塞;2) 含三通件的正常管道;3) 轻度堵塞管道,直管内设置 20 mm 模拟堵塞物;4) 重度堵塞管道,直管内设置 55 mm 模拟堵塞物,如图 5 所示为所采集 4 种运行工况下管道的时域和频域波形图。

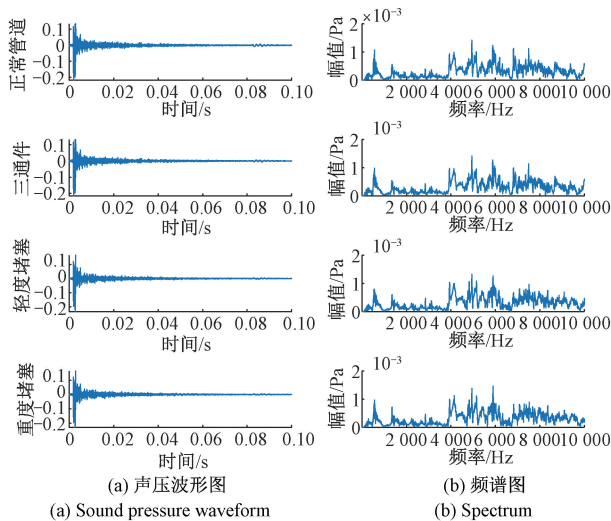


图 5 声学检测信号时域和频域波形图

Fig. 5 Acoustic waveform in time domain and frequency domain

观察时域波形图可以看出声波在管道内部的传播轨迹具有强衰减、非平稳、非线性的特点,利用管道系统过程的声压响应信号来分析其运行状况时,从原始信号难以观察到不同工况之间的差别,需要对原始信号做进一步的分析处理以实现管道状态识别的目的。

3.2 声压信号预处理和特征提取

由于测量环境以及耦合共振的影响,测量信号中存在较高频率的噪声,为提取到更有效的故障信息,采用小波分解^[19]对带噪声的原始信号进行降噪处理并选择有效频段进行特征提取。如图 6 所示,对采集的声学信号进行 6 层小波分解,图 6 中 h 为高通滤波器, g 为低通滤波器, $\downarrow 2$ 为降采样滤波器。原始声学信号经过第一层被分解成频率范围为 0~22 050 Hz 的低频分量和频率范围为 22 050~44 100 Hz 的高频分量,去除 6 000 Hz 以上高频噪声部分,依次经过 6 层分解后原始信号被分割为 8 个分量,其频率由低到高排列:0~689, 689~1 378, 1 378~2 068, 2 068~2 757, 2 757~3 446, 3 446~4 134, 4 134~4 823, 4 823~5 513 Hz。

梅尔倒谱系数可以从声音信号中提取出表征声音本质的特征信息,它能从总体上反映出管道不同运行状态的差异^[20],管道声压信号具有非平稳、非线性的特点,利用梅尔滤波器组对声压信号进行滤波和加权可获得不同频率范围内的能量分布。MFCC 的特征提取具体流程如图 7 所示,通过这些处理方式,能最大化保留声音的有效特征从而提取特征参数^[21]。

1) 尺度转换,梅尔倒谱系数是在梅尔频率 (Mel-Frequency) 域上提取出来的参数,它与频率关系为:

$$f_{mel} = 2595 \times \lg\left(1 + \frac{f}{700}\right) \quad (10)$$

式中: f_{mel} 表示的是在梅尔频率域上的频率, f 为实际频率。

2) 对声压信号分帧和加窗后的各帧信号进行 DFT 得到其频谱,然后将能量谱通过一组梅尔尺度的三角形滤波器组,最后利用对数函数对梅尔滤波后的能量谱进行压缩,得到输出的对数能量为:

$$s(m) = \ln\left(\sum_{k=0}^{N-1} |X_m(k)|^2 H_m(k)\right) \quad 0 \leq m \leq M \quad (11)$$

式中: M 为梅尔滤波器个数,通常取 22~26。

3) 将上述的对数能量代入离散余弦变换,最后得到 L 阶倒谱系数:

$$c(n) = \sqrt{\frac{2}{M}} \sum_{m=1}^N s(m) \cos\left(\frac{\pi(m-0.5)n}{M}\right) \quad (12)$$

$$n = 1, 2, \dots, L$$

式中: L 为 MFCC 特征向量的阶数。一般使用 26 个梅尔滤波器,可得到 26 个倒谱系数,取前 13 个系数作为 MFCC 特征参数,标准的 MFCC 只反映声音信号的静态

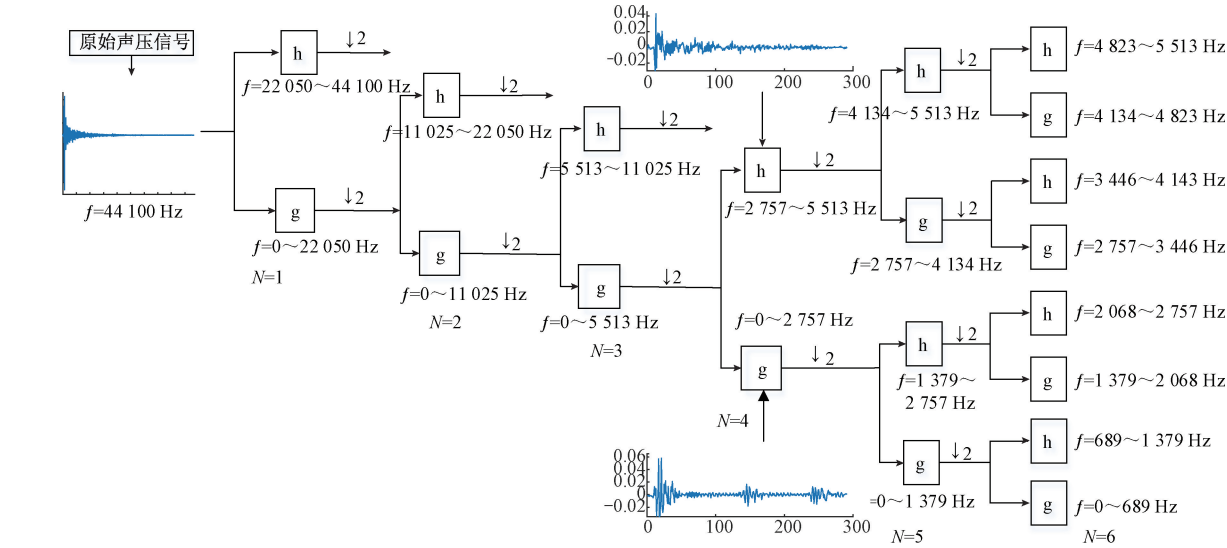


图 6 6 层小波分解图
Fig. 6 Six-level wavelet decomposition tree

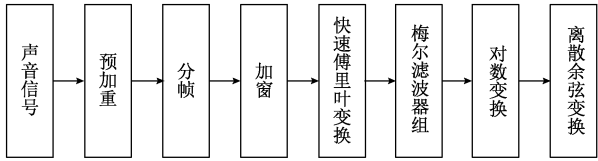


图 7 MFCC 提取流程图
Fig. 7 Flow chart of MFCC extraction

特性,可以用静态特性的差分谱来描述信号的动态特性。
分别采集正常管道、三通件管道、轻度堵塞和重度堵塞管道的声学信号,共计 200 个样本,利用小波分解对信号进行相关处理,然后提取其 26 阶 MFCC 系数,其中包含了 13 阶静态特性 MFCC 系数和 13 阶动态特性 MFCC 系数(一阶差分),部分结果如图 8 所示。

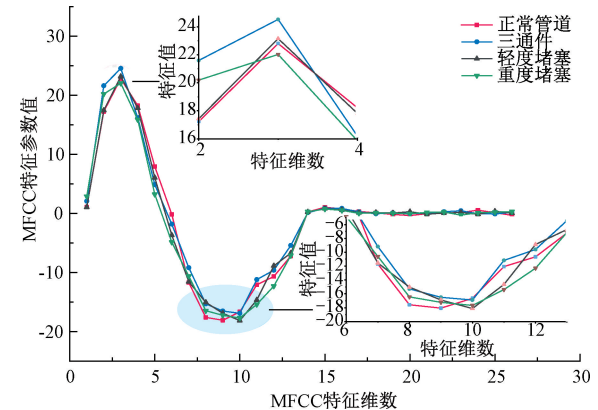


图 8 4 种工况下声信号 MFCC 图
Fig. 8 MFCC of acoustic signal under four working conditions

3.3 识别结果和算法对比分析

为验证所提模型的有效性,从管道 4 种工况的声学信号中各取 50 组样本,进行上述处理后求取每组样本的 26 个梅尔倒谱系数,构成特征数据集,标注 10%作为初始有标注样本集,剩余 90%为无标注样本集,设置第 1 批查询样本数量为 20,利用有标注数据集训练基于随机森林的分类器,计算无标注数据集中样本的信息度和相似度,得到每个样本的综合得分,如图 9(a)所示为第 1 次查询样本的综合分数信息图,选择得分最高的实例并将其从池中删除,此时不查询此实例的标签,之后重新计算未标记池中样本分数,直到达到所设置查询数量的实例后查询所有实例标签,然后扩充初始有标注样本集,所查询的 20 个实例分布如图 9(b)所示,可以看出,通过主动学习所查询未标注数据集中综合分数高的样本,有效扩充了初始故障样本的数量,使每类故障特征样本集更加丰富,从而解决有标注故障样本集较少的问题。

此时更新后的样本集为 20%有标注样本集和 80%无标注样本集,将数据集分为训练集和测试集,比例为 3:7 且样本类别均衡,首先根据欧氏距离度量指标建立特征向量集的完全连接图,利用少量有标注样本作为树的根节点构建最小生成树,其次计算未标记样本数据至根节点的最小路径,进而基于最小路径将其他未标记训练样本自动连接到对应最优路径树上实现该分类器组的构造,并对测试集进行分类识别,得到管道堵塞物以及三通件识别率。为了清晰地展示所提出方法对每类工况故障样本的识别情况,将其与基于 OPF 算法的半监督分类模型进行对比,识别结果的混淆矩阵如图 10 所示,纵坐标表示样本的实际类别,横坐标表示预测类别。当未标注

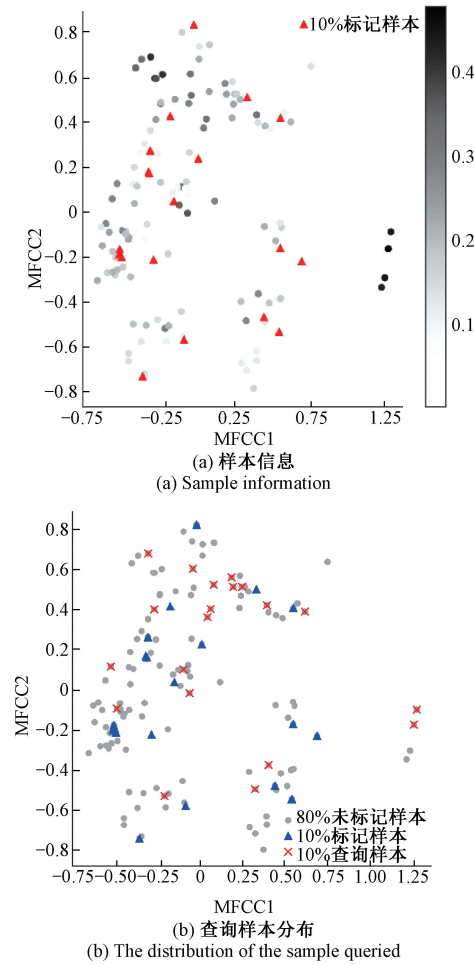


图 9 查询样本图

Fig. 9 Chart of the sample queried

样本比例为 90% 时,从图 10 可知,用基于 OPF 算法的半监督分类模型进行分类识别,正常管道和重度堵塞故障都有大量样本被错分。在图 10(a) 中,19 个正常管道样本和 6 个重度堵塞样本被错分;在图 10(b) 中,各类样本均能被准确识别,可以看出本文提出方法能提升故障识别效果。

为验证提出的模型在 10% 有标注样本集下对各类管道堵塞故障的识别效果,对采集的 4 种工况下的管道堵塞故障数据设置 5 组实验进行对比,并采用 Recall 和 F1-score 作为试验的评价指标来反映各种故障类型分类情况,Recall 能反映每类故障被正确分类的比重,F1-score 反映模型精确率和 Recall 的综合水平,其值越高表明模型越稳健。第 1 组将主动学习采样策略设置为随机采样,通过随机筛选未标注样本的方式扩充初始有标注集,然后用扩充后的数据集训练最优路径森林;第 2 组设置主动学习采样策略为分类熵采样,然后利用扩充后数据集对最优路径森林进行训练;第 3 组采用本文所提出的主动学习采样策略扩充样本集,然后训练基于 KNN 标

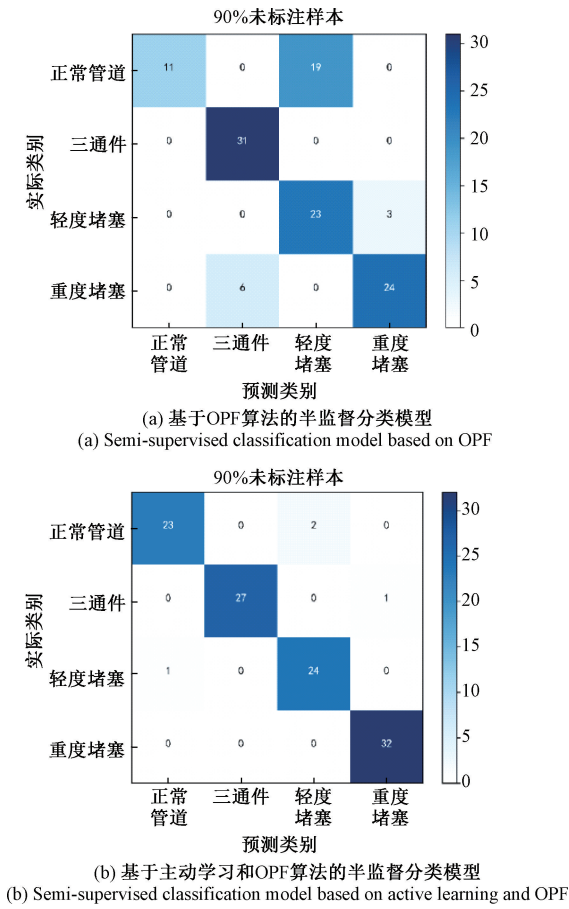


图 10 90% 未标注样本比例下的混淆矩阵图

Fig. 10 Confusion matrix with 90% unlabeled sample

签传播算法的半监督分类模型;第 4 组采用本文所提出的主动学习采样策略扩充样本集,然后训练半监督支持向量机模型;第 5 组采用本文提出方法训练最优路径森林,使用测试集验证各组试验的识别结果,对 Recall 和 F1-score 值进行对比,结果如表 1 和 2 所示。

表 1 不同主动学习采样策略识别结果

Table 1 Recognition results using different active learning sampling strategies

运行状态	Recall			F1-score		
	随机采样+OPF	分类熵+OPF	本文方法	随机采样+OPF	分类熵+OPF	本文方法
正常	0.40	0.52	0.92	0.58	0.68	0.93
三通件	0.84	0.64	0.96	0.81	0.51	0.98
轻度	0.45	0.78	0.96	0.70	0.69	0.94
重度	0.84	0.91	1.00	0.83	0.71	0.98

从表 1 可知,当采用不同的主动学习采样策略时,随机采样和分类熵采样的主动学习方法效果均不理想,表明通过简单的样本筛选方法增加有标注样本数量并不能使模型学到更多的特征信息,而本文提出的基于 BvSB 和

表 2 不同半监督算法识别结果
Table 2 Recognition results using different semi-supervised methods

运行 状态	Recall			F1-score		
	AL+KNN	AL+S3VM	本文方法	AL+KNN	AL+S3VM	本文方法
正常	0.73	0.63	0.92	0.82	0.64	0.93
三通件	0.78	0.82	0.96	0.89	0.89	0.98
轻度	0.92	0.72	0.96	0.69	0.74	0.94
重度	0.78	0.65	1.00	0.88	0.61	0.98

余弦相似度的排序批处理主动学习在只有 10%有标注数据时,Recall 值和 F1-score 值相较于其他 2 种方法都有提升,可以有效扩充各类故障数据。从表 2 可知,当采用不同的半监督算法时,AL+S3VM 方法的故障识别效果最差,重度堵塞故障的 Recall 值和 F1-score 值只有 0.65 和 0.61,而本文方法两个指标能达到 0.96 和 0.95,相较于其他两种方法都有明显提升,可以对各类故障数据进行准确识别。

为了更直观地展示本文方法在 90%未标注数据情况下与其他 4 种方法相比有更好的故障识别效果,将训练集分别输入 5 种模型,然后用训练好的模型识别测试集样本,图 11 表示不同的样本采样策略和不同半监督分类算法对分类结果的影响。

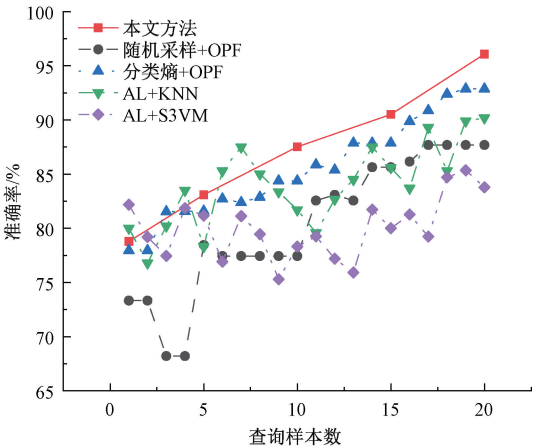


图 11 90%未标注样本比例下学习曲线图
Fig. 11 Learning curve graph under 90% unlabeled sample

从图 11 可知,本文结合 BvSB 和余弦相似度的排序批处理主动学习采样策略优于随机采样和分类熵样本采样策略,随机采样不能保证采集的样本对分类器有改善,分类熵采样在多分类问题中不能排除不重要类属概率值的影响,而最优与次优标号原则综合考虑所属概率最高的前两个类别,结合余弦相似度选择了对构建最优路径森林价值最高的样本,从而使得模型的学习准确率保持稳定上升。本文基于 OPF 算法的半监督分类模型优于基于 KNN 和 S3VM 半监督分类模型,基于 KNN 的半监

督分类模型的性能易受到近邻样本参数的影响,而 S3VM 算法在经过主动学习扩充样本集后,模型性能也没有明显提高,总体而言,本文方法在无标注样本比例例为 90%的情况下,对管道堵塞故障的识别效果优于其他 4 种方法。

3.4 时间复杂度对比

本文算法的时间复杂度由排序批处理主动学习和最优路径森林两个部分组成,其中排序批处理主动学习的时间复杂度为 $O(m^2)$,在构建最优路径森林的过程中,生成训练数据集 D 最小生成树的时间复杂度为 $O(n^2)$,计算节点路径代价值的时间复杂度为 $O(n)$,整个算法的时间复杂度为 $O(m^2+n^2+n)$,即 $O(n^2)$, m 为查询标签的数量, n 为数据集 D 的大小;随机采样和分类熵采样策略的算法复杂度均为 $O(m^2)$,因此随机采样+OPF 方法和分类熵采样+OPF 方法的时间复杂度都为 $O(n^2)$;计算 KNN 的时间复杂度为 $O(kn^2)$, k 为近邻参数,因而 AL+KNN 方法的时间复杂度为 $O(m^2+kn^2)$,即 $O(n^2)$;半监督支持向量机算法的时间复杂度为 $O(n^3)$,因此 AL+S3VM 方法的时间复杂度为 $O(n^3)$ 。

4 结 论

基于声学主动检测的管道堵塞故障诊断过程中,由于在实际运行状态下获取有标记样本的成本较高而大量包含丰富信息的无标记样本却没有得到充分利用,传统的有监督识别模型对管道的状态识别准确性和有效性会出现较大偏差,本文提出了一种基于主动学习和最优路径森林的故障分类识别方法,首先利用主动学习筛选价值高的样本批量更新训练样本集,然后通过构建最优路径森林来学习样本特征空间的底层结构,实现对管道堵塞状态的识别分类。最后用实验室采集的管道工况数据集验证了所提出方法的有效性,实验结果表明,主动学习方法通过采样策略筛选出选最具有信息量的样本更新模型而没有充分利用剩余无标注样本中包含的信息,而半监督学习可以提取无标注样本的全局结构信息,有助于建立更加精准的故障分类模型,实现这两类方法的有机组合使用,最大程度的改善了分类模型的性能,同时提高了无标记数据的利用率。通过排序批处理的主动学习采样方式,根据每个样本的信息量和相似度采集样本,能在减少算法本身耗时的同时有效避免采集到重复、无意义的样本,通过主动学习引入的真值也能稳定提升半监督分类模型的性能。基于最优路径森林算法的半监督分类模型相对于其他基于距离度量提取训练样本全局结构信息的半监督分类模型可以达到更高的故障识别率,且模型参数少,易于训练,能够有效识别管道的故障状态,具有广泛的适用性。

由于管道运行环境的复杂性和堵塞形式的多样性,使得对于管道的堵塞问题还需进一步探究,下一步工作将针对运行中管道存在的不同形态的堵塞和损伤故障设计实验内容,扩充样本集数量和管道状态类别进一步完善分类模型,同时对声波在不同介质环境下的管道内的传播特性、衰减特性和能量场分布规律展开机理分析,为实验研究和工程应用提供理论依据和指导。

参考文献

- [1] CAMARGO G, BUGATTI P H, SAITO P T M. Active semi-supervised learning for biological data classification[J]. Plos One, 2020, 15(8): e0237428.
- [2] JIAN C X, YANG K J, AO Y H. Industrial fault diagnosis based on active learning and semi-supervised learning using small training set [J]. Engineering Applications of Artificial Intelligence, 2021, 104.
- [3] 韩彦岭, 赵耀, 周汝雁, 等. 协同主动学习和半监督方法的海冰图像分类[J]. 海洋学报, 2020, 42(1): 123-135.
HAN Y L, ZHAO Y, ZHOU R Y, et al. Cooperative active learning and semi-supervised method for sea ice image classification[J]. Haiyang Xuebao, 2020, 42(1): 123-135.
- [4] HAN W J, COUTINHO E, RUAN H B, et al. Semi-supervised active learning for sound classification in hybrid learning environments [J]. Plos One, 2016, 11(9): e0162075.
- [5] YIN L L, WANG H G, FAN W H, et al. Incorporate active learning to semi-supervised industrial fault classification[J]. Journal of Process Control, 2019, 78: 88-97.
- [6] 周乾, 曾周末, 黄新敬, 等. 基于主动声学的管道球形内检测器示踪定位方法[J]. 仪器仪表学报, 2020, 41(8): 217-225.
ZHOU Q, ZENG ZH M, HUANG X J, et al. Tracing and localization method of spherical inner detector for pipelines based on active acoustics[J]. Chinese Journal of Scientific Instrument, 2020, 41(8): 217-225.
- [7] DATTA S, SARKAR S. A review on different pipeline fault detection methods[J]. Journal of Loss Prevention in the Process Industries, 2016, 41: 97-106.
- [8] 邓勇, 曹敏, 赖治屹. 基于深度学习的天然气管道气体压力超声检测模式识别方法[J]. 电子测量与仪器学报, 2021, 35(10): 176-183.
DENG Y, CAO M, LAI ZH L. Ultrasonic detection pattern recognition method for natural gas pipeline gas pressure based on deep learning [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(10): 176-183.
- [9] 焦敬品, 李勇强, 吴斌, 等. 基于 BP 神经网络的管道泄漏声信号识别方法研究[J]. 仪器仪表学报, 2016, 37(11): 2588-2596.
JIAO J P, LI Y Q, WU B, et al. Research on acoustic signal recognition method for pipeline leakage with BP neural network [J]. Chinese Journal of Scientific Instrument, 2016, 37(11): 2588-2596.
- [10] 赵燕锋, 冯早, 朱雪峰, 等. 复杂工况下基于时频图像和 CNN-SVM 的管道堵塞识别研究[J]. 电子测量与仪器学报, 2021, 35(2): 161-170.
ZHAO Y F, FENG Z, ZHU X F, et al. Research on blocking recognition of drainage pipeline under complicated conditions based on time frequency image and CNN-SVM [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(2): 161-170.
- [11] 龙军, 殷建平, 祝恩, 等. 主动学习研究综述[J]. 计算机研究与发展, 2008(S1): 300-304.
LONG J, YIN J P, ZHU EN, et al. A survey of active learning [J]. Journal of Computer Research and Development, 2008(S1): 300-304.
- [12] 王显龙, 冯早, 朱雪峰, 等. 基于改进的 QBC 和随机森林的管道类别非均衡堵塞故障识别[J]. 电子测量与仪器学报, 2021, 35(3): 84-93.
WANG X L, FENG Z, ZHU X F, et al. Blockage recognition method of drainage pipeline learning from unbalanced data based on improved QBC and random forest [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(3): 84-93.
- [13] CARDOSO T N C, SILVA R M, CANUTO S, et al. Ranked batch-mode active learning [J]. Information Sciences, 2017, 379: 313-337.
- [14] PAPA J P, FALCAO A X, SUZUKI C T N, et al. A discrete approach for supervised pattern recognition[C]. International Workshop on Combinatorial Image Analysis, Springer, Berlin, Heidelberg, 2008: 136-147.
- [15] AMORIM W P, FALCAO A X, CARVALHO M H, et al. Semi-supervised pattern classification using optimum-path forest [C]. 2014 27TH Sibgrapi Conference on Graphics, Patterns and Images (SIBGRAPI), 2014: 111-118.
- [16] AMORIM W P, FALCAO A X, PAPA J P, et al. Improving semi-supervised learning through optimum connectivity[J]. Pattern Recognition, 2016, 60: 72-85.
- [17] LI J, ZHU Q. Semi-supervised self-training method based on an optimum-path forest [J]. IEEE Access, 2019, 7: 36388-36399.
- [18] 伍林峰, 冯早, 朱雪峰. 基于优化 VMD 和连续隐马

尔科夫模型的管道堵塞状态评估[J]. 振动与冲击, 2020, 39(22): 214-222, 233.

WU L F, FENG Z, ZHU X F. Pipeline blocking state assessment based on optimized VMD and continuous hidden Markov model [J]. Journal of Vibration and Shock, 2020, 39(22): 214-222, 233.

- [19] 李志农, 朱明, 褚福磊, 等. 基于经验小波变换的机械故障诊断方法研究[J]. 仪器仪表学报, 2014, 35(11): 2423-2432.

LI ZH N, ZHU M, CHU F L, et al. Mechanical fault diagnosis method based on empirical wavelet transform[J]. Chinese Journal of Scientific Instrument, 2014, 35(11): 2423-2432.

- [20] SHANTHAKUMAR S, SHAKILA S, SUNETH P, et al. Environmental sound classification using deep learning[J]. Instrumentation, 2020, 7(3): 15-22.

- [21] 刘思思, 谭建平, 易子馗. 基于MFCC和SVM的车窗电机异常噪声辨识方法研究[J]. 振动与冲击, 2017, 36(5): 102-107.

LIU S S, TAN J P, YI Z K. A window motor abnormal noise identification method based on MFCC and SVM[J]. Journal of Vibration and Shock, 2017, 36(5): 102-107.

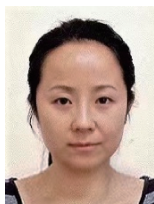
作者简介



李恬, 2020年毕业于周口师范学院获得学士学位, 现为昆明理工大学硕士研究生, 主要研究方向为信号处理、机器学习算法。

E-mail: 1364381330@qq.com

Li Tian received her B. Sc. degree from Zhoukou Normal University in 2020. Now she is a M. Sc. candidate at Kunming University of Science and Technology. Her main research interests include signal process and machine learning algorithm.



冯早(通信作者), 2009年于英国纽卡斯尔大学获得硕士学位, 2014年于英国布拉德福德大学获工学博士学位, 现为昆明理工大学副教授, 主要研究方向为声学无损检测、数据挖掘、机器学习算法研究。

E-mail: 6483975@qq.com

Feng Zao(Corresponding author) received her M. Sc. from the University of Newcastle in 2009 and Ph. D. from University of Bradford in 2014. Now she is an associate professor at Kunming University of Science and Technology. Her main research interests include acoustic nondestructive testing technology, application research, data mining and machine learning algorithm based on acoustics.