DOI: 10.13382/j.jemi.B2003343

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*

王显龙[1,2] 冯 早[1,2] 朱雪峰[1,2] 赵燕锋[1,2]
(1. 昆明理工大学 信息工程与自动化学院 昆明 650500; 2. 云南省人工智能重点实验室 昆明 650500)

摘 要:针对城市埋地排水管道堵塞故障检测过程中有标签故障样本少,管道运行状态样本集存在类别不均衡及样本标注成本高昂的问题,提出一种基于主动学习的排水管道堵塞故障分类识别方法。该方法采用改进后的委员会样本查询策略通过基于一致熵的委员会样本查询策略建立主动学习模型来实现不均衡样本集的学习。经过充分考虑样本的信息度并挖掘信息度高的未标注样本进行标注后,结合多个随机森林分类器组成委员会对未标注样本进行分类识别。在实验室所采集的管道运行数据集上对委员会样本查询策略中的投票熵、一致熵和随机选择样本查询策略进行对比验证。实验结果表明,采用基于一致熵的委员会查询策略在类别分布均衡初始训练集下有更快的收敛速度和更好的稳定性,在类别非均衡分布的初始训练集下同样具有良好的识别效果。

关键词: 埋地管道;类别不均衡;委员会样本查询;随机森林;一致熵
中图分类号: TN06;TP274.2 **文献标识码**: A **国家标准学科分类代码**: 510.40

Blockage recognition method of drainage pipeline learning from unbalanced data based on improved QBC and random forest

Wang Xianlong[1,2] Feng Zao[1,2] Zhu Xuefeng[1,2] Zhao Yanfeng[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 University of Science and Technology, Kunming 650500, China)

Abstract: Aiming at the problems of few labeled fault samples, unbalanced dataset of pipeline operation state data set and high cost of sample labeling in the process of urban buried drainage pipeline blockage fault detection, an classification and recognition method of drainage pipeline blockage fault based on active learning is proposed. This method adopts the improved committee sample query strategy, and established an active learning model based on consensus entropy to realize the learning of unbalanced data set. After fully considering the uncertainty of the samples and mining the most informative of unlabeled samples for labeling, the committee composed of several random forest classifiers was used to classify and identify the unlabeled samples. The vote entropy, uniform entropy and randomly selected sample query strategy are compared and verified on the pipeline operation data set collected by the laboratory. The experimental results show that the committee query strategy based on consensus entropy has faster convergence speed and better stability under the initial training set of class distribution equilibrium, and also has good recognition effect under the initial training set with unbalanced distribution of categories.

Keywords: buried pipeline; category imbalance; query-by-committee; random forest; consensus entropy

0 引 言

随着城市的发展,地下排水管网已成为重要的基础设施。但因管道服役时间的不断延长管道会出现老化、断裂等各种现象,管道易发生堵塞事故[1]。排水管道堵塞容易造成严重的污水外溢,特别是汛期,由于在降水不断增加的过程中,雨量的集中会导致严重的内涝,影响城市的整体排水能力,严重威胁到城市居民的生活和居住安全,因此如何准确判断管道的安全运行具有重要的意

收稿日期: 2020-07-18 Received Date: 2020-07-18
*基金项目:国家自然科学基金(61563024,51765022)项目资助

义[2]。目前针对城市埋地排水管道的主要检测方法有基于 CCTV 管道闭路电视系统和管道潜望镜检测技术等,然而这些方法存在检测滞后、成本高昂等问题[3]。近年来,由于声波检测的成本低廉、设备安装简便和优良的检测效果等优点被广泛用于管道结构缺陷和功能缺陷的检测。通过对携带管道运行状态的声波信号进行相应的预处理和特征提取可将管道堵塞检测问题转化为基于数据驱动的识别问题。

根据文献[2]实地调研情况来看管道的运行状态数据集具有明显的类别不均衡问题。在管道的故障识别过程中,正常管道样本的数量远大于各类故障样本的数量,在样本分布上表现出明显的类别间不均衡特性。若将正常管道样本错判为堵塞故障,由此将产生不必要的人工成本;若将堵塞管道样本错判为正常管道,将会造成更为严重的后果,由于错判带来的故障排除的延误有可能对城市交通,生产和民生带来较大的影响。在解决不均衡数据集分类问题时,传统的分类算法以正确率为主要优化目标导致少数类故障样本数据识别率不够理想。解决数据集类别不均衡问题在数据层面主要有欠采样和过采样两种方法。如 SMOTE 过采样算法通过对少数类数据合成来增加少数类样本数量使得数据样本类别均衡,但 SMOTE 算法由于增加了大量样本导致其模型训练时间过长,且 SMOTE 算法易受到样本中噪声信息影响从而降低分类性能[4];郎宪明等[5]通过利用 K 均值聚类欠采样方法降低管道泄漏数据集中的不平衡比例,最大限度保留多数类样本原始分布信息从而提高了管道泄漏的识别率,但 K 均值聚类欠采样方法中聚类数量的确定极大影响了分类效果。在算法层面的改进主要是对赋予少数类更大的惩罚系数提高对少数类样本分类的正确率,何大伟等[6]通过代价敏感支持向量机对轴承的内圈故障和正常样本进行分类识别,但该方法并不适用于多类别的不均衡数据集分类。基于数据层面的处理方法由于改变了数据的初始分布情况有可能降低分类器的性能;基于算法层面的方法关键在于错分代价的选取,大多数情况下的错分代价是基于经验选取,在面对多类别数据时很难选取出一个合适的错分代价。

近年来主动学习在故障诊断领域得到越来越广泛的应用,相较于监督学习,主动学习大大降低了人工标注样本的成本。根据未标记样本采样策略的不同可将采样策略分为基于不确定度的采样策略、基于泛化误差减小的采样策略、基于版本空间缩减的采样策略[7]。其中基于版本空间缩减采样策略的代表性方法是 QBC 委员会投票算法[8]。该方法主要利用初始已标注样本训练两个或多个分类器并构建“委员会”,委员会评估未标注样本集并选择评估结果差别最大样本来交由人工标注。其中衡量委员会成员之间的差异的标准为投票熵,投票熵值越大表明该样本在委员会间产生了较大的分歧,然而投票熵并没有考虑未标记样本类属概率值容易导致漏选重要的样本并导致学习的准确率不稳定。同时在面对不均衡数据集时,唐明珠等[9]建立基于改进的投票委员会选择和代价敏感支持向量机的模型,提高了铜闪速炉熔炼过程中工业故障的识别准确率。徐海龙等[10]采用投票熵和 KL 散度的委员会采样策略同时结合多个改进的代价敏感支持向量机,降低了样本不均衡对分类结果的影响。上述两种方法均采用代价敏感支持向量机的方法,该方法虽然提高了学习效率优化了分类结果,但是忽略了采用基于委员会的主动学习策略需要确保委员会成员间的高品质及成员间的差异的原则,并且代价敏感支持向量机更适用于二分类问题,并不适用于管道堵塞故障识别中的多分类问题。针对管道检测过程中数据不均衡分布的特性和综合考虑上述方法的缺点后,本文提出一种基于一致熵的委员会采样策略和随机森林的主动学习方法对管道堵塞进行分类识别。

1　管道声学原理分析

若检测声波的波长远大于管道的直径,则管道中声波以平面波的形式传播,而堵塞物可被视为声学负载,导致管内的平面波声场受到干扰[11]。若管道内部有堵塞物,堵塞物的法向表面声阻抗为 Z_a。主动激励声源所发射的声波会受到堵塞物的干扰形成反射声波,另外一部分声波被堵塞物吸收,还有一部分声波会绕过堵塞物发生衍射。假设管内激励声源的入射声波为 P_i,经过堵塞物的反射波为 P_r,入射声波绕过管道堵塞物产生衍射波 P_t,3 种声波的公式表示为:

$$\begin{cases} P_i = P_{ai} e^{j(\omega t - kx)} \\ P_r = P_{ar} e^{j(\omega t + kx)} \\ P_t = P_{at} e^{j(\omega t - kx)} \end{cases} \tag{1}$$

由于声波在管道内部传播遇到堵塞物受限于声压连续、体积速度连续两种声学边界条件。声压连续的表示了入射波、反射波和衍射波之间的联系,公式表示为:

$$P_{ai} + P_{ar} = P_{at} \tag{2}$$

声波在含三通件构件的管道的传播时,管内的入射波 P_i 经过阻抗为 Z_b 的三通件时在主管产生反射波 P_r 和透射波 P_t,在与三通件连接的旁支中产生漏入波 P_b。根据声压连续原则和声学边界条件可得含三通件构件的声波声压关系式为:

$$P_i + P_r = P_l = P_b \tag{3}$$

体积速度关系式为:

$$\frac{s}{\rho_0 c_0}(P_i - P_r) = \frac{sP_l}{\rho_0 c_0} + \frac{P_b}{Z_b} \tag{4}$$

声波的传递过程实质是声波能量的传播过程,声波在管道中传播使管道内媒质产生扰动,媒质被迫产生压缩和膨胀[12]。管道中声场的能量又两部分组成一部分是管道内介质震动产生的动能 ΔE_k 和由于受到声波影响产生了形变势能 ΔE_p。

$$\Delta E_k = \frac{1}{2}(\rho_0 V_0) v^2 \tag{5}$$

$$\Delta E_p = -\int_0^P p\mathrm{d}V \tag{6}$$

单位体积 V_0 的声能量和为:

$$\Delta E = \Delta E_k + \Delta E_p = \frac{V_0}{2}\rho_0\left(v^2 + \frac{1}{\rho_0^2 v_0^2}p^2\right) \tag{7}$$

2 信号采集与特征提取

2.1 管道运行数据采集

实验数据来自英国布拉德福德大学管道实验室[13]。实验装置包括 4 个水听器、扬声器、声卡、功率放大器和 WinMLS 测试软件。为模拟排水管道工作状况,实验管道采用长 15.4 m、直径 150 mm 的黏土管道。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射到管道内部。扬声器发出声波信号后,4 通道的水听器采集反射声波通过采集卡和滤波器与计算机相连接,实验采样频率设置为 44 100 Hz。

为模拟排水管道的真实运行状况,管道内部保持在管道中间部位距离管道口和扬声器 7 m 位置放置堵塞物来模拟管道堵塞,其中定义堵塞物高度若超过管道直径的 1/3 则为中重度堵塞,堵塞物高度不超过管径的 1/3 则为轻度堵塞。为排除管道构件三通件对实验结果的影响,实验设计了 4 种工况分别如下:1)直管管内无堵塞,管内保持较低的水位;2)轻微堵塞,直管内设置 20 mm 模拟堵塞物;3)中重度堵塞,直管内设置 55 mm 模拟堵塞物;4)管道中段位置设置三通件构件,管内保持较低水位。为模拟实际检测过程中的数据间类别分布不均衡的特性,实验共采集 305 组实验信号其中无堵塞直管样本个数为 115,含三通件构件管道样本数为 115,轻度堵塞管道样本数为 42,重度堵塞样本数为 33。

2.2 完全集合经验模态分解(CEEMDAN)

CEEMDAN[14] 是在集合经验模态分解(ensemble empirical mode decomposition, EEMD)[15] 分解过程的每一阶段加入自适应白噪声使其分解过程即具完整性又能抑制噪声的一种算法[16],其计算步骤如下。

1)设定 $\omega(n)$ 为高斯白噪声,$E_k(\cdot)$ 为对信号通过经验模态分解(EMD)产生的第 k 个模态分量,CEEMDAN 所产生的第 k 个模态分量记为 $\widetilde{IMF}_k$。

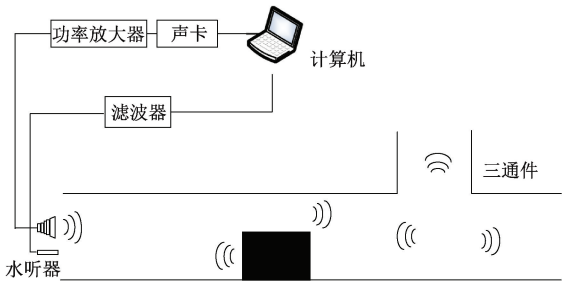


图 1 实验模拟管道运行工况示意图

Fig. 1 Schematic diagram of pipeline operation condition

2)对信号 $s(n)$ 进行 EMD 分解得到 $IMF_k(n)$,k 为 IMF 分量个数,n 为信号的采样点个数。定义 $\overline{IMF_k}$ 为由 EEMD 分解得到的均值,则该均值就是第 1 个 CEEMDAN 分解的分量记为 $\tilde{C}_{IMF_1(n)}$。

3)计算第 1 个剩余分量:

$$r_1(n) = s(n) - \tilde{C}_{IMF_1(n)} \tag{8}$$

4)计算信号 $S(n)$ 的第 2 个 IMF 分量:

$$\tilde{C}_{IMF_2(n)} = \frac{1}{n}\sum_{m=1}^{n} E_1\{r_1(n) + \varepsilon_1 E_1[\omega(n)]\} \tag{9}$$

由此计算出第 k 个剩余分量为:

$$r_k = r_{k-1}(n) - \tilde{C}_I MF_k(n) \tag{10}$$

5)计算第 $k+1$ 个 IMF 分量:

$$\tilde{C}_{IMF_{k+1}(n)} = \frac{1}{n}\sum_{m=1}^{1} E_1\{r_k(n) + \varepsilon_k E_k[\omega(n)]\} \tag{11}$$

6)执行步骤 5)直到获取的余量信号不能被分解为止,信号最终被分解为:

$$s(n) = \sum_{k=1}^{k} \tilde{C}_{IMF_k} + R(n) \tag{12}$$

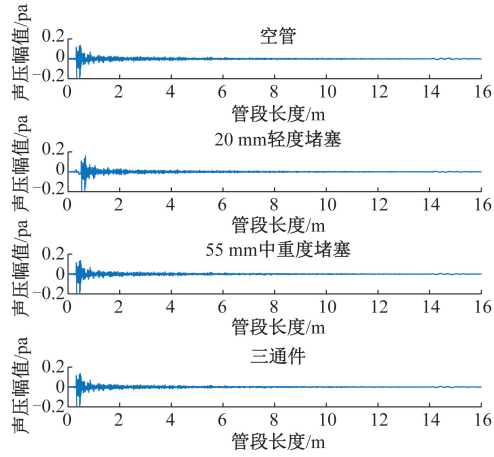


图 2 四种工况时域波形

Fig. 2 Time domain waveform of original signal

从实验中采集的正常直管、含三通件构件管道、轻度堵塞管道、和重度堵塞管道 4 种管道运行工况的时域波

形如图 2 所示。观察时域信号波形图可以看出主动激励声波在管道内部的信号特性具有强衰减、非平稳、非线性的特点,仅从原始信号难以观察出不同工况之间的差别。通过对原始信号进行 CEEMDAN 分解可有效地将管道堵塞物和三通件造成管道内部信号分解为本征模态分量 IMF,因此对所采集信号的进行进一步的特征提取就更为重要。对采集得到的信号进行 CEEMDAN 分解,得到 12 个 IMF 分量。图 3 所示为原始空管信号经 CEEMDAN 分解后得到的 12 个 IMF 分量。其中 $IMF_1 \sim IMF_{12}$ 分别表示有高频到低频的不同频段成分。

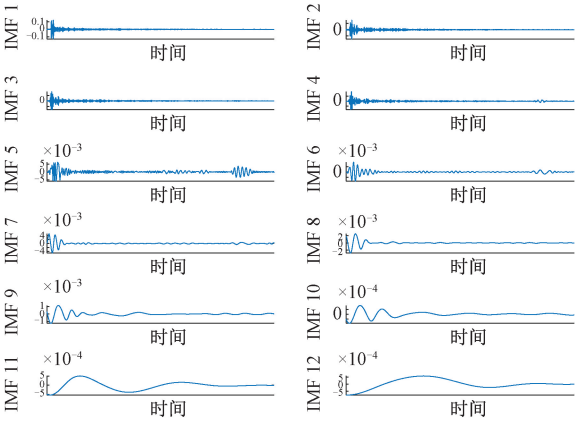


图 3　空管原始信号的 CEEMDAN 分解

Fig. 3　CEEMDAN decomposition of original normal signal

2.3　IMF 分量选取和特征提取——平均声能量密度

若将 CEEMDAN 分解后得到的 12 的 IMF 分量都进行特征提取会造成特征冗余,增加计算成本。本文通过计算各个 IMF 分量和原始信号的相关系数来确定有效 IMF 分量,将各个 IMF 分量与原始信号的相关系数如图 4 所示。若 IMF 分量的相关系数越大说明和原始信号之间相似性越大。根据参考文献[17],本文选取相关系数大于 0.1 的 IMF1 ~ IMF6 分量并进一步提取各个分量的平特征。

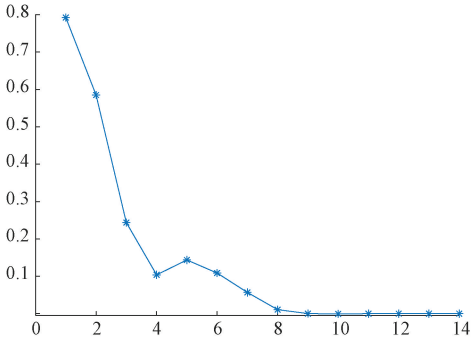


图 4　IMF 分量的相关系数

Fig. 4　Correlation coefficient of IMF component

由于声强表示声波在介质内沿传播方向在单位时间内、单位面积所做的功,其计算公式为:

$$I = \frac{1}{T}\int_0^T Re(P)Re(v)\,\mathrm{d}t \tag{13}$$

将式(1)和(2)代入式(14)可得沿声波传播方向的声波的声强为:

$$I_i = \frac{1}{2}\rho_0 c_0 v_a^2 \tag{14}$$

若管道中存在堵塞则反射声波的声强为:

$$I_r = -\frac{1}{2}\rho_0 c_0 v_a^2 \tag{15}$$

式中:ρ_0 为声场介质密度;c_0 为声速;V_a 为管内媒质质点本身速度。当管内的检测声波遇到堵塞物时产生反射声波时,这时管内的总声强为 $I = I_i + I_r$。如果入射声波和反射声波相等总声强为零,管内声场中声强这一物理量往往不能反映其能量关系。

为了提取能够有效表征管内声场能量变化的特征,根据式(7)计算出管道内部声波信号的声能量单位时间平均值为:

$$\overline{\Delta E} = \frac{1}{T}\int_0^T \Delta E\,\mathrm{d}t = \frac{1}{2}V_0\frac{P_a^2}{\rho_0 c_0^2} \tag{16}$$

则声能量密度 ε 为:

$$\varepsilon = \frac{\Delta E}{V_0} = \frac{1}{2}\rho_0\left(v^2 + \frac{1}{\rho_0^2 v_0^2}p^2\right) \tag{17}$$

则单位体积 V_0 的平均声能量密度 $\bar{\varepsilon}$ 表达式为:

$$\bar{\varepsilon} = \frac{\overline{\Delta E}}{V_0} = \frac{p_a^2}{2\rho_0 c_0^2} \tag{18}$$

而平均声能量密度这一物理量能够反映管道内部声场存在反射波时能量的变化。其各个工况信号的 $IMF_1 \sim IMF_6$ 的平均声能量密度如表 1 所示。

表 1　平均声能量密度特征

Table 1　Feature of average acoustic energy density

模态分量	各类信号平均声能量密度			
	正常管道	含三通件管道	轻度堵塞管道	重度堵塞管道
IMF1	13.506 0	14.177 7	14.388 3	13.682 0
IMF2	13.576 7	11.765 6	11.837 8	11.613 0
IMF3	12.848 8	13.156 8	13.339 3	13.261 0
IMF4	4.958 6	3.851 7	4.735 7	5.256 6
IMF5	2.689 9	2.629 2	3.241 0	3.601 3
IMF6	4.836 9	5.364 3	3.075 1	1.117 9

3　基于一致熵和随机森林的管道堵塞故障检测

3.1　主动学习

采用监督学习方法建立管道堵塞故障分类模型时,

大量已标注样本是分类模型获得良好的分类精度的必要条件。实际上,不同数据样本对于学习模型的贡献度是不一样的,如果能够选取一部分最有价值的数据进行标注,有可能仅基于少量数据就能获得同样高效的模型。为了实现这一目标,关键在于如何选择出最有价值的数据样本并去获取它们的标记信息。主动学习就是研究这一问题的一种机器学习框架。其核心任务是制定选择样本的标准,从而选择尽可能少的样本进行标注来训练出一个好的学习模型[18]。针对管道检测过程中故障样本较少导致管道数据集类别不均衡和样本标注过程需要耗费大量的代价的问题,本文基于一致熵和随机森林的主动学习来检测识别管道堵塞故障,其具体步骤如下。

1)采集管道信号并提取声学信号特征。

2)标注部分管道数据样本作为初始标注训练集 L,其余未标注的管道数据样本作为未标注样本集 U,选取两个随机森林构建委员会,设置未标注样本查询选择次数 T。

3)利用已标注的管道数据样本集 L 训练委员会中的基分类器——随机森林。

4)委员会评估未标注样本集 U,依据式(22)计算未标注样本集 U 中的一致熵,根绝一致熵的大小对未标注样本进行排序,并选择一致熵最大的样本交给专家标注。

5)更新已标注训练集 L 和未标注样本集 U。

6)判断委员会的分类精度是否达到标准和迭代次数是否达到限制。若满足分类精度要求输出分类识别结果,若不满足分类精度要求循环步骤 3)和 4)。

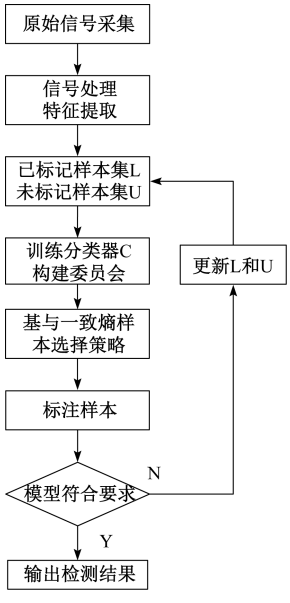


图 5 主动学习流程

Fig. 5 Flow chart of active learning

3.2 改进的 QBC 样本查询策略——一致熵

在基于委员会的样本查询策略中,分类器将收到一系列未标记的示例作为输入。为此,根据当前训练集的统计数据构造由两个或多个分类器组成的“委员会”。然后,每个委员会成员对候选示例进行分类判断,学习者测量委员会成员之间的分歧程度,并为每个示例决定是否要为其标记。委员会查询样本选择策略算法是基于版本空间缩减采样策略的著名算法[19],其中衡量委员会对未标记样本的分歧程度主要有投票熵,投票熵的定义为:

$$Vote_{entropy} = \operatorname{argmax} \frac{1}{\log\min(k, |C|)} \sum_{i=1}^{C} \frac{V(C,e_i)}{k} \log \frac{V(C,e_i)}{k} \tag{19}$$

式中:K 表示委员会中委员会成员的个数,$V(C,e_i)$ 表示分类器对样本 X_i 分类为 C_j 类的票数。当投票熵的值越大,委员会对该样本的分歧程度也就越大,该样本就越容易导致分类错误影响分类效果。然而,投票熵的缺点是它没有考虑委员会成员的分类概率分布可能会导致错失信息度丰富的样例。

考虑到投票熵的缺点,本文提出一致熵样本查询策略。一致熵衡量了未标注样本类属概率的不稳定程度,其算法定义由式(20)给出。与计算选票分布不同一致熵首先计算每个分类器下未标注样本的类属概率值,将每个未标注样本的类属概率值相加除以委员会中分类器的个数求平均值,该概率平均值称为一致概率,在此基础上计算每个样本的一致熵概率并选取具有最大一致熵的样本。该算法的具体步骤如下。

定义基准分类器个数为 t,数据集中类别数为 C,有 n 个未标注的数据样本。

1)计算每个基准分类器对每个样本的类属概率记为 $P_{ij}^{k}(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant C, 1 \leqslant k \leqslant t)$ 得到 t 个 $n \times C$ 概率矩阵,每个概率矩阵如下:

$$\begin{matrix} P_{11}^{1} & \cdots & P_{1C}^{1} & P_{11}^{k} & \cdots & P_{1C}^{k} & P_{11}^{t} & \cdots & P_{1C}^{t} \\ \vdots & \ddots & \vdots & \cdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ P_{n1}^{1} & \cdots & P_{nC}^{1} & P_{n1}^{k} & \cdots & P_{nC}^{k} & P_{n1}^{t} & \cdots & P_{nC}^{t} \end{matrix} \tag{20}$$

2)将这 t 个概率矩阵的对应位置的元素相加得到每个分类器的概率平均值矩阵,其中矩阵的每行代表一个样本,每个元素代表样本的类属概率值。

$$\begin{matrix} \overline{P_{11}} & \cdots & \overline{P_{1C}} \\ \vdots & \ddots & \vdots \\ \overline{P_{n1}} & \cdots & \overline{P_{nC}} \end{matrix} \tag{21}$$

3)步骤 2)得到的概率矩阵每一行表示一个样本的类属概率分布,类比熵的概念计算一致熵:

$$Consensus_{entropy} = \operatorname{argmax} \frac{1}{\alpha} \sum_{n=1}^{c} \overline{P_{ij}} \log(\overline{P_{ij}}) \tag{22}$$

3.3 随机森林构建委员会

决策树是一种树状分类器,在树的每个节点通过选

择最优的分裂特征进行分类到达停止条件时停止,每个决策树的子节点代表不同类别的数据。决策树的一般生成方式为从上而下的生成方式,一般来说一颗决策树包含一个根节点、若干个内部节点和若干个叶节点,叶节点对应决策结果,其他每个节点对应于一个属性测试;根节点表示所有样本集合,从根节点到叶节点的路径表示区分不同类别的待测样本。通常,决策树是一种简单快速且高效准确的分类方法,当面对高维复杂数据时决策树分类性能提升有限;此外,决策树可能会对样本集特征空间过度划分导致过拟合问题发生。

随机森林是由多棵并行的互不相关的决策树构建的集成学习模型,随机森林解决了决策树性能瓶颈问题,和决策树模型相比具有更好的泛化能力、不易产生过拟合、对异常值不敏感等优点[20]。因此本文在构建委员会时选择两个随机森林分类模型作为评审委员会。

4　结果与分析

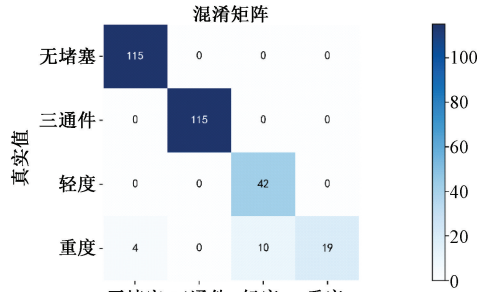
由于实际过程中的排水管道堵塞故障数据集具有类间分布不均衡特性,本文为模拟了实际工况下采集获取的管道样本类别比例,选取空管无堵塞样本为 115 组,含三通件空管样本 115 组,轻度堵塞样本数 42 组,中重度堵塞样本 33 组。为验证本文提出的基于一致熵和随机森林的主动学习模型的性能设置了两组实验:设置实验 1 使用类别均衡的初始训练集,分别采用一致熵采样策略、投票熵采样策略和随机采样策略的主动学习方法,对比不同采样策略对学习模型最终识别效果的影响。随机选取已标注初始训练集样本数 8 组,每个类别样本数各两组。实验 2 为了进一步验证本位所提方法的性能,对比在类别不均衡分布初始训练集下不同采样策略对识别结果的影响,其中选取正常空管类样本 4 组,含三通件空管样本 4 组,轻度堵塞和中重度堵塞样本数各一组。两组实验的主动查询次数均设置为 20 次,两组实验的初始训练集样本分布如表 2 所示。经过 20 次样本迭代选择后,实验 1 利用 28 个已标注训练样本对剩下 277 个未标注样本集测试,其分类结果的混淆矩阵如图 6 所示;实验 2 利用 30 个已标注训练样本对剩下 275 个未标注样本集测试,其分类结果的混淆矩阵如图 7 所示。

表 2　两组实验的初始训练集样本分布

Table 2　Sample distribution of initial training set for two groups of experiments

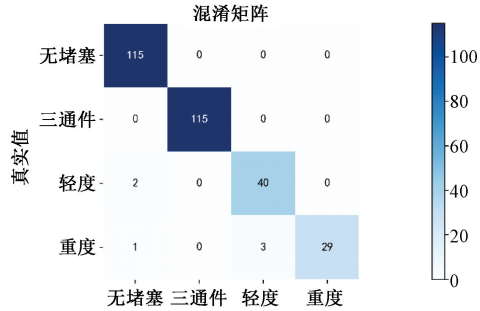
	标注训练集样本数	初始训练集各类样本数比例
实验 1	8	1∶1∶1∶1
实验 2	10	4∶4∶1∶1

图 6(a)为采用随机采样策略最终学习效果的混淆矩阵,从混淆矩阵可以看出,38%的中重度堵塞样本被错分类为轻度堵塞,12%的中重度堵塞样本被错分类为空管,说明采用随机采样策略不能有效降低数据集类别不均衡对分类效果的影响。图 6(b)为采用投票熵采样策略的最终学习效果的混淆矩阵,其中 5%的轻度堵塞样本被错分类为空管,3%的中重度堵塞样本被错分类为空管,另有 9%的中重度堵塞样本被分类为轻度堵塞,其余样本均识别正确。投票熵样本查询策略考虑了样本的不确定性,在一定程度上改善了少数类样本的分类效果但是稳定性较差。图 6(c)为采用一致熵的采样策略的学习效果混淆矩阵,由于一致熵从未标注样本的类属概率角度衡量未标注样本的不确定性,在标注样本时提供了



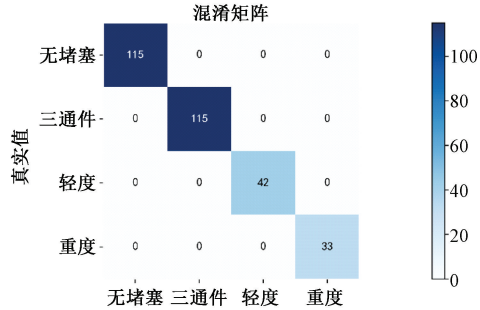
(a) 基于随机采样的混淆矩阵

(a) Confusion matrix based on random sampling



(b) 基于投票熵采样的混淆矩阵

(b) Confusion matrix based on vote entropy sampling

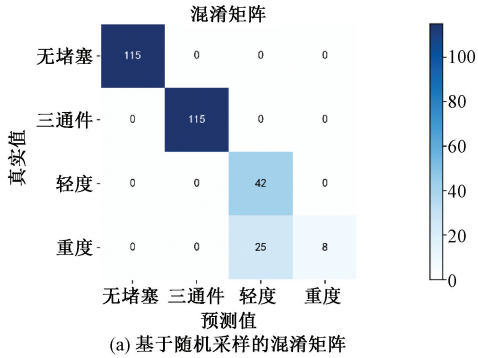


(c) 基于一致熵采样的混淆矩阵

(c) Confusion matrix based on consensus entropy sampling

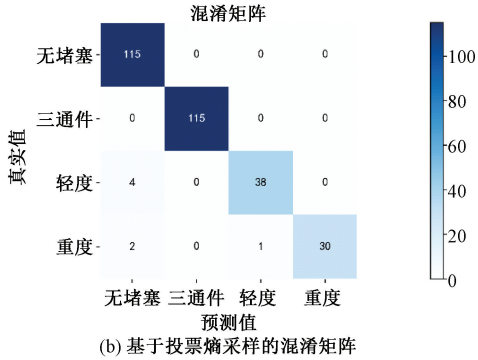
图 6　均衡训练集下 3 种采样策略的混淆矩阵

Fig. 6　Confusion matrix of three sampling strategies under balanced training set



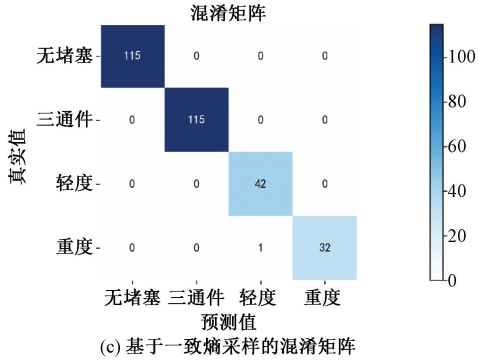
(a) 基于随机采样的混淆矩阵

(a) Confusion matrix based on random sampling



(b) 基于投票熵采样的混淆矩阵

(b) Confusion matrix based on vote entropy sampling



(c) 基于一致熵采样的混淆矩阵

(c) Confusion matrix based on consensus entropy sampling

图 7 非均衡训练集下 3 种采样策略的混淆矩阵

Fig. 7 Confusion matrix of three sampling strategies under unbalanced training set

信息度最大的样本,在面对少数类样本分类时取得了较高且稳定的准确率。

从图 7(a)可以看出,采用随机采样策略的主动学习模型,76%的中重度堵塞样本被错分为轻度堵塞;如图 7(b)为采用投票熵采样策略的主动学习,将 10%的轻度堵塞样本错分为空管,3%的中重度堵塞错分为轻度堵塞,6%的中重度堵塞被错分为空管;采用一致熵采样策略的主动学习模型识别结果如图 7(c)所示,仅有 3%的中重度堵塞样本被错分为轻度堵塞,其余样本均识别正确。实验结果对比得出:在相同的标注次数下,采用一致熵的采样策略可以有效消除数据不均衡分布给分类结果造成的干扰并减少人工标注负担,对未标注样本的分类准确率明显高于采用其他两种采样策略,且模型性能也在前 5 次迭代之后明显趋于稳定。

4.1 实验 1 结果分析

图 8 所示为初始训练集为类别均衡的情况下学习准确率随着迭代次数的变化。从图 8 的训练效果来看 3 种采样策略均取得了 90%以上的准确率,在主动学习的过程中随着样本迭代查询次数的增加,有标签样本集的扩展使分类性能明显提升。基于一致熵的样本采样策略取得了最佳的学习效果,随着迭代次数的不断增加分类模型的性能保持了较稳定的状态。采用基于随机采样的样本查询策略由于没有考虑未标注样本的不确定性导致分类器性能较不稳定,分类效果波动较大。采用基于投票熵的样本查询策略的分类效果好于随机采样效果,但和一致熵相比由于投票熵没有充分考虑样本的概率分布导致在样本查询时选择未标注样本有可能选择到孤点,导致稳定性不如一致熵采样策略,准确率有小幅度的波动。

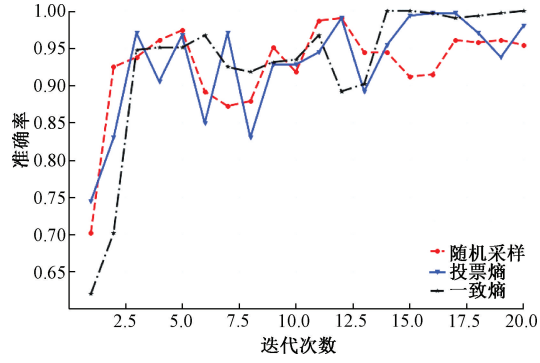


图 8 均衡训练集下 3 种采样策略的学习率对比

Fig. 8 Learning rate of 3 sampling strategies under balanced training sets

4.2 实验 2 结果分析

实验 2 设置已标注初始训练集为类别不均衡分布,各类别样本数量比例设置为 4∶4∶1∶1,原因是正常管道以及包含三通配件的正常管道在管道服役期间是主体存在,而存在堵塞故障的管道数量应远小于正常管道。如图 9 所示 3 种样本采样策略前 5 次迭代由于受到初始训练数据集类别非均衡分布的影响导致学习准确率均出现了较大的波动,随着主动学习迭代次数的增加,各采样策略的准确率均有了较大的提升。采用随机采样策略的主动学习模型的准确率为 91.8%。采用投票熵采样策略的主动学习模型其学习准确率达到了 97.7%,但是准确率有较大的波动模型稳定性较差。采用一致熵采样策略的主动学习模型在前 5 次迭代查询时经历了准确率下降,但随着初始训练集扩展,模型的分类效果取得了明显的提升,最终达到 99.7%的准确率。

比较图 8 和 9 的学习曲线可以看出,若初始训练集

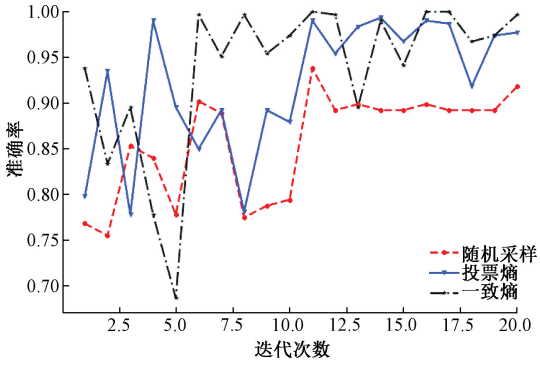


图 9　非均衡训练集下 3 种采样策略的学习率对比

Fig. 9　Learning rate of 3 sampling strategies under unbalanced training sets

为不均衡的样本集,模型的准确率会出现波动,这是由于委员会在迭代的早期缺乏少数类样本的先验知识容易出现误分类现象。随着样本查询次数的增加,已标注样本集逐渐扩展,从图 6(a)和 7(a)可以看出基于随机选择策略的主动学习模型分类性能较差并不能保证模型的稳定性;从图 6(b)和 7(b)可以看出基于投票熵的委员会主动学习模型准确率虽然有所提升,但是模型稳定性还是受到少数类样本数量的影响;从图 6(c)和 7(c)的可以看出基于本文提出的一致熵的委员会主动学习方法仅在实验 2 中有一个中重度堵塞样本被误分类为轻度堵塞,在数据不均衡的情况下模型也表现出了良好的稳定性。

5　结　论

针对排水管道进行故障排查时经常会面临标注管道样本数据代价昂贵耗时,以及现实情况下管道堵塞故障的数据类型分布不均衡的问题,本文提出基于一致熵和随机森林的排水管道堵塞识别方法。实验结果得出以下结论:

1)利用声学主动信号检测管道堵塞是一种有效的检测手段,平均声能量密度能够有效表征管内声场由于受到堵塞影响的信号变化。

2)针对现实情况下已标注样本量较少的问题本文建立基于主动学习的管道堵塞识别分类模型,通过 20 次的迭代查询并标注有效地节省了标注成本。

3)针对类别间样本分布不均衡的问题,本文通过改进投票熵算法提出一致熵算法结合委员会查询策略从未标注样本的类属概率的角度出发,通过衡量委员会成员间的差异选出最大分歧度的样本,该采样策略方法充分考虑了每个分类器在面对数据类别不均衡分布情况下的分类差异性,极大提升了模型的分类效果,同时模型在面对类别均衡或不均衡初始训练集时表现出了更强的稳定性,准确率更高。

由于该模型所使用的数据样本均为实验环境下所采集的信号,信号受环境等噪音干扰较小,下一步的工作重点是引入噪声点和离群点数据以及相应的评价机制排除噪声数据的干扰,进一步提升分类模型的泛化能力和综合性能。

参考文献

[1] 申陈俊. 城市排水管道堵塞及疏通技术研究[J]. 华东科技(学术版),2016(12):343-343.
SHEN CH J. Study on blockage and dredging technology of urban drainage pipeline [J]. East China Science and Technology(Academic Edition),2016(12):343-343.

[2] 罗东旭. 城市河流沿线排水管道检测及修复技术研究[J]. 冶金丛刊,2019,4(6):97-99.
LUO D X. Research on detection and repair technology of drainage pipeline along urban river [J]. Metallurgical series, 2019, 4(6):97-99.

[3] 李洋,冯早,黄国勇,等. 基于广义 Fisher-互信息的管道堵塞故障特征选择方法[J]. 电子测量与仪器学报,2018,32(11):1-8.
LI Y, FENG Z, HUANG G Y, et al. Feature selection method for pipe blockage based ongeneralized Fisher-mutual information [J]. Journal of Electronic Measurement and Intrumentation,2018, 32(11):1-8.

[4] BLAGUS R, LUSA L. SMOTE for high-dimensional class-imbalanced data. [J]. BMC Bioinformatics, 2013 14(1): 106. .

[5] 郎宪明,李平,曹江涛,等. 基于非平衡数据处理的管道泄漏检测与定位研究[J]. 湖南大学学报:自然科学版, 2018, 45(2):115-123.
LANG X M,LI P,CAO J T,et al. Study on pipeline leak detection and location based on imbalanced data processing [J]. Journal of Hunan University: Natural Science Edition, 2018, 45(2):115-123.

[6] 何大伟,彭靖波,胡金海,等. 基于改进 FOA 优化的 CS-SVM 轴承故障诊断研究[J]. 振动与冲击,2018,37(18):113-119.
HE D W, PENG J B, HU J H, et al. Bearing fault diagnosis based on a modified CS-SVM model optimized by an improved FOA algorithm [J]. Journal of Vibration & Shock, 2018, 37(18): 113-119.

[7] 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.

[8] 赵悦,穆志纯. 基于委员会投票选择方法的主动学习的研究[J]. 太原理工大学学报,2006(4): 469-472.
ZHAO Y, MU ZH CH. Rsearch on query-by-committee method of active learning [J]. Journal of Taiyuan University of Technology, 2006(4): 469-472.

[9] 唐明珠,阳春华,桂卫华. 基于改进的 QBC 和 CS-SVM 的故障检测[J]. 控制与决策, 2012, 27(10): 1489-1493.
TANG M ZH, YANG CH H, GUI W H. Fault detection based on modified QBC and CS_SVM [J]. Control and Decision, 2012, 27(10):1489-1493.

[10] 徐海龙,别晓峰,冯卉,等. 一种基于 QBC 的 SVM 主动学习算法[J]. 系工程与电子技术, 2015, 37(12): 2865-2871.
XU H L, BIE X F, FENG H, et al. Active learning algorithm for SVM based on QBC [J]. Systems Engineering and Electronics, 2015, 37(12):2865-2871.

[11] 马大猷. 现代声学理论基础[M]. 北京:科学出版社, 2004.
MA D Q. Theoretical basis of Modern Acoustics [M]. Beijing: Science Press2004

[12] 伍林峰,冯早,黄国勇,等. 小波包增强稀疏表征分类的管道堵塞故障识别[J]. 电子测量与仪器学报, 2019,33(3):35-43.
WU L F, FENG Z, HUANG G Y, et al. Pipeline jam fault identification based on wavelet packet enhanced sparse representation classification [J]. Journal of Electronic Measurement and Instrumentation, 2019, 33(3):35-43.

[13] ZAO F. Condition classification in underground pipes based on acoustical characteristics [D]. Bradford: University of Bradford,2013.

[14] 耿读艳,王晨旭,赵杰,等. 基于 CEEMDAN-PE 的心冲击信号降噪方法研究[J]. 仪器仪表学报, 2019, 40(6): 155-161.
GENG D Y, WANG CH X, ZHAO J, et al, Research on BCG signal de-noising method based on CEEMDAN and PE [J]. Chinese Journal of Scientific Instrument, 2019, 40(6):155-161.

[15] 韩庆阳,孙强,王晓东,等. CEEMDAN 去噪在拉曼光谱中的应用研究[J]. 激光与光电子学进展, 2015, 52(11):274-280.
HAN Q Y, SUN Q, WANG X D, et al. Application of CEEMDAN in Raman spectroscopy denoising [J]. Laser & Optoelectronic Progress, 2015, 52 (11): 274-280.

[16] 陈仁祥,汤宝平,吕中亮. 基于相关系数的 EEMD 转子振动信号降噪方法[J]振动. 测试与诊断, 2012, 32(4): 542-546,685.
CHEN R X, TANG B P, LV ZH L. Ensemble emperical mode decomposition de-noising method based on correlation coefficients for vibration signal of rotor system [J]. Vibration Test and Diagnosis, 2012, 32 (4): 542-546,685.

[17] 古莹奎,曾磊,张敏,等. 基于 CEEMDAN-SQI-SVD 的齿轮箱局部故障特征提取[J]. 仪器仪表学报,2019, 40(5):78-88.
GU Y K, ZENG L, ZHANG M, et al. Feature extraction method for garbox local fault based on CEEMDAN-SQI-SVD [J]. Chinese Journal of Scientific Instrument, 2019,40(5):78-88.

[18] 叶晨,王宏志,高宏, 等. 面向众包数据清洗的主动学习技术[J]. 软件学报,2020,31(4):1162-1172.
YE CH, WANG H ZH, GAO H, et al. Active learning approach for crowdsourcing-enhanced data cleaning [J]. Journal of Software, 2020,31 (4):1162-1172.

[19] 杨文柱,田潇潇,王思乐,等. 主动学习算法研究进展[J]. 河北大学学报(自然科学版),2017,37(2): 216-224.
YANG W Z, TIAN X X, WANG S L, et al. Advances in active learning algorithms [J]. Journal of Hebei University(Natural Science Edition), 2017, 37(2): 216-224.

[20] 王丽婷,丁晓青,方驰. 基于随机森林的人脸关键点精确定位方法[J]. 清华大学学报(自然科学版), 2009, 49(4):85-88.
WANG L T, DING X Q, FANG CH. Accurate localizationg of facial feature points based on random forest classifier [J]. Journal of Tsinghua University (Science & Technology)2009,49(4):85-88.

作者简介

王显龙, 2018 年于南阳师范学院获得学士学位,现为昆明理工大学硕士研究生,主要研究方向为机器学习和故障诊断。
E-mail:1021604005@qq.com

Wang Xianlong received his B. Sc. degree in 2018 from Nanyang Normal University. Now he is a M. Sc. candidate in Kunming University of Science and Technology. His main research interests include fault diagnosis and machine learning.

冯早，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得博士学位，现为昆明理工大学副教授，主要研究方向为基于声学的无损检测技术应用及研究、数据挖掘、机器学习算法研究。

E-mail：6483975@ qq. com

Feng Zao received M. Sc. from University of Newcastle upon Tyne in 2009, and Ph. D. from University of Bradford in 2014. And now she is an associate professor at Kunming University of Science and Technology. Her main research interests include non-destructive testing technology application research, data mining and machine learning algorithm based on acoustics.