DOI: 10.13382/j.jemi.B1902765

基于 FOA 优化的 CSSVM 管道堵塞状态识别研究*

王 菲[1,2] 冯 早[1,2] 朱雪峰[1,2]

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

摘 要:针对城市排水管道正常与堵塞故障状态在数据获取上的不平衡性造成的运行状态识别准确率下降的问题,提出了一种基于果蝇优化算法的代价敏感支持向量机的管道堵塞状态识别方法。根据排水管道内各运行状态下采集到的不平衡数据集,首先对不平衡数据集进行小波包分解,其次,提取各个分解系数的能量熵、近似熵指标构建特征向量集合;采用果蝇优化算法(FOA)对不同类样本惩罚因子 C_m 和核函数参数 g 进行优化选取,即对代价敏感支持向量机(CS-SVM)模型优化,将特征集合输入优化后的 CS-SVM 模型中,对排水管道的正常和堵塞状态识别,通过增大对少数类样本错分的惩罚代价,结果表明,提升了少数类的识别准确率。

关键词: 管道堵塞;果蝇优化算法;代价敏感支持向量机

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

Research on CSSVM pipe jam status recognition based on FOA optimization

Wang Fei[1,2] Feng Zao[1,2] Zhu Xuefeng[1,2]

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

Abstract: Aiming at the problem of the accuracy of the recognition of the operating state caused by the unbalanced data acquisition in the normal and blocked fault state of the drainage pipeline, a method for a pipeline clogging state recognition based on cost-sensitive support vector machine based on fruit fly optimization algorithm is proposed. According to the unbalanced data set collected under various operating conditions in the drainage pipeline, the wavelet packet decomposition is first performed on the unbalanced data set. Secondly, the energy entropy of each decomposition coefficient and the approximate entropy index are used to construct the feature vector set. The fruit fly optimization algorithm is adopted. (FOA) optimizes the penalty factor Cm and the kernel function parameter g, that is, the cost-sensitive support vector machine (CS-SVM) model optimization, and inputs the feature set into the optimized CS-SVM model to normalize the drainage pipe. Blocking state recognition, by increasing the penalty cost of misclassification of a few types of samples, the results show that the recognition accuracy of a few classes is improved.

Keywords: pipeline blockage; fly optimization algorithm; cost-sensitive-support-vector-machine

0 引 言

城市排水网络在多雨季进行径流并在一定程度上保护城市免受洪水的影响,复杂的城市总体规划和基础设施公用事业的开发导致排水系统交叉冗杂,使用年限较长的管道逐渐出现腐蚀、泄漏、堵塞等故障,排水管道出现堵塞点后因为水流面的逐渐减小,堵塞程度逐渐增强,未及时发现,极易导致排水管道堵塞,甚至引发人类安全隐患,所以,及时检测管道堵塞,控制堵塞程度量变造成的危害,对我国城市的稳健发展具有重要意义[1]。

近年来,国内外研究人员的研究热点着眼于将人工智能算法应用于管道运行状态的监测与识别中,通过机器学习可以自主学习获得管道运行状态监测与识别的模

收稿日期:2019-11-19 Received Date: 2019-11-19

*基金项目:国家自然科学基金(61563024)资助项目

型,提取有效特征参数后输入模型进行模式识别,常用的信号预处理方法有经验模态分解、变分模态分解及小波变换等。用于模式识别的方法主要有 BP 神经网络、支持向量机(SVM)等,此类模型对平衡数据集具有较强的泛化能力,而体现该类算法泛化能力的前提条件是选取特征适当和较大的数据量,多数基于统计学习的分类算法对训练集的样本分布有较严格的均衡性要求,将机器学习分类算法应用于具有不平衡特征的数据集时,存在一些弊端,可能会造成极端值、数据样本稀缺及噪声等一系列问题,限制了训练得到的分类器的性能,对故障类的识别效果欠佳,整体分类性能不理想[2]。

目前,此类不平衡数据集广泛存在于实际应用中,如在检测某种特定疾病艾滋病、乳腺癌等中,患该疾病的样本数量远远小于检查总体人数;网站防御中,网络被入侵或中毒的机率也远远小于正常访问的机率;在过滤垃圾邮件中,垃圾邮件的数据样本量远小于正常邮件的数据样本量等,闫慈等[3]提出了采用不同 4 种重采样算法,以代谢综合征为例,研究不平衡数据对分类算法的影响;许玉格等[4]提出一种基于加权极限学习机集成算法的污水处理故障诊断建模方法,提高了故障类别的识别率;谭洁帆等[5],提出了一种采用 Triplet-sampling 的卷积神经网络和代价敏感支持向量机(CS-SVM)的不平衡图像分类方法—Triplet-CSSVM,提高了少数类查全率、使分类结果总代价降低。

针对较小数据样本量支持向量机对其具有较好的学习能力,因为遵循结构风险最小化原则,文玉梅等[6]使用一对一 SVM 多分类方法,并对管道泄漏信号进行了有效识别;康守强等[7]以 SVM 为基础,提出了基于果蝇优化算法-多样支持向量机(FOA-MKSVM)的滚动轴承故障分类方法,研究了管道故障小样本模式识别方法,提高了管道故障的识别精度。但在样本类不平衡的问题中,SVM 分类算法会在多类及类间数据样本中产生数据样本冗杂重复、过拟合等问题[8],为了降低故障识别错误代价,许多代价敏感学习算法均已在故障诊断领域逐渐被提出。因此基于 SVM 的代价敏感学习算法应用于故障识别等研究具有非常重要的实际意义。刘永斌等[9]研究了基于 SVM 的代价敏感故障诊断,较好地解决了二分类故障识别问题。但是应用于解决多分类故障问题时均存在困难,李艳霞等[10]总结了通过分析训练集的先验信息,通过 SVM 将类间不平衡的样本设置不相同的惩罚系数,验证了代价敏感学习方法的可行性;何大伟等[11]优化了代价敏感支持向量的多个惩罚因子参数,对公开数据集 IMS 航空轴承数据进行识别,能够有效处理误分类代价不同的轴承故障问题。

针对城市排水管道正常与堵塞故障状态在数据获取上的不平衡性造成的运行状态识别准确率下降的问题,本文旨在决策层面解决不平衡数据集的分类识别问题,根据排水管道内各运行状态下采集到的不平衡数据集,首先对不平衡数据集进行小波包分解,其次,提取各个分解系数的能量熵、近似熵指标构建特征向量集合;采用 FOA 对不同类样本惩罚因子 C_m 和核函数参数 g 进行优化选取,即对 CS-SVM 模型优化,对排水管道的正常和堵塞状态识别,通过增大对少数类样本错分的惩罚代价,提升对少数类的识别准确率。

1　CS-SVM 及其参数

1.1　SVM

SVM 最早出现在 Vapnik 针对规模有限的样本数量下,针对机器学习的研究因为其理论研究的限制,并未形成完整的理论,随着科技时代的发展,基于较为成熟的统计学习理论,逐渐形成了 SVM 机器学习整体框架[12]。SVM 的基本求解原理即为针对方程中最优超平面分类问题,如图 1 所示。

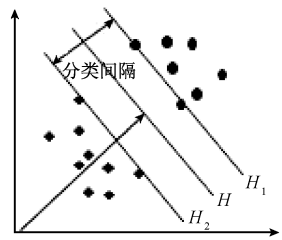


图 1　分类原理

Fig. 1　Classification schematic

设定样本集向量$\{x_i\}$,$(i=1,2,\cdots,n)$根据固定的非线性映射,映射后集合为无限维特征空间 Z,即映射条件为 $\varphi:R^m\to Z, x\to\varphi(x)$,即最优超平面 $H:\boldsymbol{w}\cdot\{\varphi(x)\}+b=0$ 将训练样本集准确地分为两类。

$$\begin{aligned} H_1&:\boldsymbol{\omega}\cdot\{\varphi(x)\}+b=1-\xi_i \\ H_2&:\boldsymbol{\omega}\cdot\{\varphi(x)\}+b=-1+\xi_i \end{aligned} \tag{1}$$

式中:H_1、H_2 表示为两个平行于 H 的超平面,两个平面之间因无样本点且距离 H 最远,此最远距离称为分类间隔,增加了 ξ_i 定义为松弛量,其计算值为 $2/\|\boldsymbol{w}\|$;$\xi_i\geqslant 0$。

为使两超平面之间的分类间隔最大且无样本点,则约束方程描述为:

$$\begin{aligned} &y_i(\{\boldsymbol{w}\}\{\varphi(x_i)\}+b)-1+\xi_i \\ &i=1,\cdots,n \end{aligned} \tag{2}$$

1.2　CS-SVM

由式(1)、(2)联立求解最优分类超平面 H 为:

$$\min_{\{w\},b,\xi}\left[\frac{1}{2}\|\boldsymbol{\omega}\|^2+\frac{1}{2}C\sum_{i=1}^{n}\xi_i\right]=\min_{\{w\},b,\xi}\left[\frac{1}{2}\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}+\frac{1}{2}C\sum_{i=1}^{n}\xi_i\right]$$

$$\text{s.t. } y_i(\{w\}\{\varphi(x_i)\}+b)-1+\xi_i\geqslant 0 \tag{3}$$

式中:C 为惩罚因子,即为权衡对样本的拟合能力与对测试样本的预测能力。

引入 Lagrange 乘子,将约束方程转化为拉格朗日函数,求解上述凸二次规划方程,将式中 $\{\boldsymbol{\omega}\}$、b、ξ_i 等参数求偏导并设定为 0,可推导得出 Lagrange 函数的目标函数。

对于 $\xi_i=0$ 的标准支持向量,有:

$$(\{w\}\{\varphi(x_i)\}+b)=1 \tag{4}$$

即满足式(5)。

$$y_i\left[\sum_{j=1}^{n}y_j\alpha_jK(\{x_i\},\{x_j\})+b\right]=1 \tag{5}$$

对于每个标准支持向量 x_i 都有:

$$b=y_i-\sum_{j=1}^{n}y_j\alpha_jK(\{x_i\},\{x_j\}) \tag{6}$$

则满足稳定性的阈值为:

$$b=\frac{1}{I}\sum_{i\in I}\left[y_i-\sum_{j=1}^{n}y_j\alpha_jK(\{x_i\},\{x_j\})\right] \tag{7}$$

考虑到 SVM 分类算法会在类间数据样本中产生数据样本冗杂重复、过拟合等问题,即不同类数据样本采用相同惩罚因子,会产生错判,导致识别准确率降低。

在实际工况中所得数据集为不平衡时,即正常类样本数量远多于故障类样本,因此对正常类与故障类样本分别采用不同的惩罚因子 C_m 和 C_h,SVM 允许对不同的训练错误设定不同的惩罚参数来达到对样本点的不同容忍程度,即对应的最优分类超平面 H 优化问题可用公式表达为:

$$\min\left[\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\left[C_m\sum_{\{i|y_i=m\}}\xi_i+C_h\sum_{\{i|y_i=h\}}\xi_i\right]\right] \tag{8}$$

$$\text{s.t. } y_i[\boldsymbol{\omega}^{\mathrm{T}}x_i+b]\geqslant 1-\xi_i;i=1,\cdots,n$$

CS-SVM 模型公式由式(4)~(8)联立即可得到。式(8)中,为了构建多分类器模型时,权重主要偏向于数量少的故障类样本,即提高故障类样本的惩罚因子数值,其中,不同类别的惩罚因子 C_m 和 C_h 的比值关系通常是根据领域知识得到的。其中,参数选择存在人为误差,多采用优化算法例粒子群优化算法、遗传算法等对分类器模型参数进行优化,以期得到更佳的识别准确率。

2 基于 FOA 的 CS-SVM 寻优

2.1 FOA

在 2012 年 Pan[13] 提出的果蝇优化算法是一种新型的依据群体本能对参数进行优化的算法。其本质是果蝇依据食物源的位置凭借嗅觉本能搜索食物源的过程。比较于其他优化算法,例如粒子群优化算法、遗传算法等,果蝇算法中初始参数设置较少、对个体领域的搜索能力强。成功应用于优化支持向量机参数、神经网络参数等。

果蝇优化算法的基本流程可归纳以下步骤,如图 2 所示。

1)果蝇群体位置初始化:随机定义果蝇位置、种群规模 N、最大搜索步长。

2)嗅觉搜索:在果蝇种群中心(X,Y)附近,随机产生多个果蝇个体,计算个体与坐标原点的距离,计算气味浓度值,得到个体适应度。

3)选择适应度最佳的果蝇个体,保留适应度最佳果蝇个体位置,得到新的种群中心(X',Y')。

4)判断是否达到终止条件,如最大搜索步长、最佳停滞步长,若达到,输出最终解;否则,继续迭代寻优。

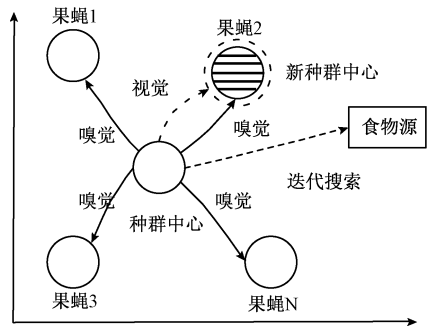


图 2 FOA 搜索过程示意图

Fig. 2 Schematic diagram of the FOA search process

2.2 FOA-CS-SVM 模型

1)初始化 FOA 参数。包括果蝇种群规模 N、迭代次数,果蝇个体初始坐标。X、Y 分别取 3 个随机数,得到果蝇种群中的初始坐标(X_0^1,Y_0^1)、(X_0^2,Y_0^2)、(X_0^3,Y_0^3)。

2)设定随机方向与距离的果蝇个体,其不同坐标为(X_i^1,Y_i^1)、(X_i^2,Y_i^2)、(X_i^3,Y_i^3)。味道浓度判定值 S_0^1、S_0^2、S_0^3 为果蝇个体坐标与原点坐标之间距离的倒数。

3)由味道浓度判定值确定 CS-SVM 初始给定参数各类数据样本惩罚因子 C_i、核参数 g 和核函数权重值 λ 的范围。$C_i=100S_i^1$,$g_i=10S_i^2$,$\lambda_i=0.1S_i^3$,即 $C\in(0,1\ 000]$,$g\in(0,100]$,$\lambda\in[0,1]$。

4)将数据归一化后的训练样本用于训练 CS-SVM 分类器模型,其中,适应度函数依据分类准确率,即适应度函数公式为:

$$S_i=Fitness(C_i,g_i,\lambda_i)=accuracy(C_i,g_i,\lambda_i) \tag{9}$$

5)寻找分类准确率最高即适应度函数最大数值所对应的果蝇个体位置,进入迭代寻优过程,并判断此时分类准确率是否大于前一初始模型最大分类准确率。若大

于,则记录最大分类准确率最大值及对应的坐标,更新果蝇种群位置坐标,即并将保留坐标重新赋值给初始坐标 (X_0^1, Y_0^1)、(X_0^2, Y_0^2)、(X_0^3, Y_0^3)。若小于,即未达到最大迭代次数,然后返回到步骤 2)。

6)记录多类不同的惩罚因子 C_m、核参数 g 和核函数 λ,建立 CS-SVM 模型。

3　基于 FOA 的 CSSVM 管道堵塞状态识别方法

针对城市排水管道正常与堵塞故障状态在数据获取上的不平衡性造成的运行状态识别准确率下降的问题,本文提出了一种基于 FOA 的 CS-SVM 管道堵塞状态识别方法。具体步骤如下。

1)信号预处理。对不平衡数据集声学信号进行 3 层小波包分解,其中,选取“db4”小波基函数。

2)特征分量。对特征分量即第 3 层的 8 个小波包分解系数进行能量分布分析。能量分布越高说明包含信号的特征信息越丰富,故而选取能量分布较高的小波包分解系数。

3)对特征小波分量分别提取能量熵、近似熵两个指标联合构建特征向量集。

4)故障识别。将所得特征向量集按比例分为训练集和测试集,将训练集输入经过果蝇算法参数优化的 CSSVM 模型中,得到 FOA-CSSVM 模型,识别测试集,得到分类识别效果。

本文方法流程如图 3 所示。

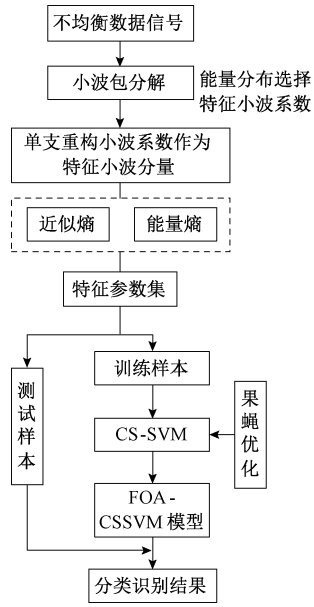


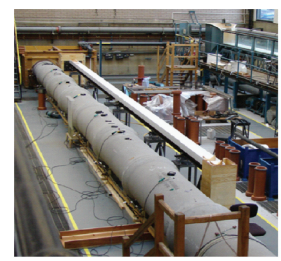
图 3　管道堵塞识别方法流程

Fig. 3　Blockage recognition method flowchart

4　数据采集与处理

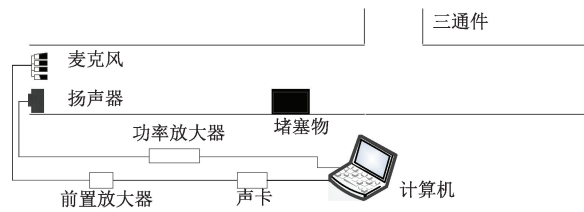
4.1　实验设计

英国布拉德福德大学管道实验室的管道堵塞实验平台所用管道为一根长 14.4 m、直径 150 mm 的黏土质管道[14]。管道左端同一竖直面上下安装有扬声器和麦克风。其中,为了减小声波能量的无效损失,管道右端放置挡板,。研究人员使用装有 Windmills 软件的计算机控制声卡产生一个 10 s 的正弦扫频声波信号,该声波信号频率范围为 100~6 000 Hz,声卡产生的声波信号经功率放大器放大后,驱动扬声器,将声波传播到管道中。麦克风接收反射回来的回波信号,因收集到的回波信号较弱,经功率放大器放大后上传至计算机存储,以便进行后续信号的处理与分析,其中管道内部放置石质挡板模拟堵塞物,轻微堵塞状态与重度堵塞状态分别用高度为 20、55 mm 的堵塞物进行模拟,其管道正常与堵塞状态检测平台的结构如图 4 所示。



(a) 管道实验平台

(a) Experimental platform structure



(b) 声学检测过程示意

(b) Schematic of acoustic detection process

图 4　实验平台结构示意图

Fig. 4　Experimental platform structure

4.2　数据预处理

实验数据分析的采样频率为 44 100 Hz,采样时间 0.1 s,实验室分别获取正常健康管道(200 组样本)、含有三通件的健康管道(160 组样本)、20 mm 轻微堵塞状态管道(60 组样本)、55 mm 重度堵塞状态管道(40 组样本)4 种状态下的声学信号,样本总量 540 组。

1)小波包分解

在对非线性非平稳信号的分析中,小波包分解具有

较为突出的分析能力和良好的时频特性[15]。

小波包分解的时频划分如图 5 所示。小波包分解是一个树形分解结构,其分解次数为 N($2N=n$,n 表示采样点数)。以 3 层分解为例,原始信号经过 1 层小波包分解,低通滤波得到低频信号,高通滤波得到高频信号,再对得到的信号分解,以此类推,最后由多个短小但信息切割较均匀的频率波段构成原完整频率信号。

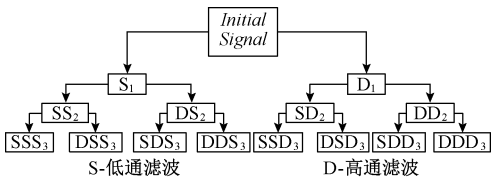


图 5　小波包分解的时频划分

Fig. 5　Time-frequency division of wavelet packet decomposition

小波包分解算法。由 $\{d_n^{j-1,m}\}$ 求 $\{d_n^{j-1,2m}\}$ 与 $\{d_n^{j-1,2m+1}\}$:

$$d_n^{j,2m+1} = \sum_k g_{k-2n} d_k^{j-1,m} \tag{10}$$

小波包重构算法。由 $\{d_n^{j,m}\}$ 求 $\{d_n^{j-1,2m}\}$ 与 $\{d_n^{j-1,2m+1}\}$:

$$d_n^{j-1,m} = \sum_k \bar{h}_{n-2k} d_k^{j,2m} + \bar{g}_{n-2k} d_k^{j,2m+1} \tag{11}$$

其中 $d_n^{j,m}$ 是序列 $f(n)$ 经过 j 层小波包分解后所得到的第 m 个分解序列,$\bar{h}$ 与 $\bar{g}$ 分别对应 h 与 g 的对偶算子。

在实际应用中,小波基于分解层数的选择影响小波包函数的去噪效果。一般来讲,db 小波系在声音去噪中是经常会被用到的一族小波基;不同检测信号、不同信噪比下,小波包函数分解层数中都存在去噪效果最好或接近最好的分解层数。

以 1 组正常健康管道信号为例,即对 0.1 s 的数据进行信号分析,如图 6 所示。

进而计算各个小波包系数的能量占比。得到的能量占比结果如图 7 所示。由图 7 直方图与折线图可以看出正常管道信号第 5 个和第 6 个小波包分解系数能量占比数值较小,其和整体 4 类显现出的规律相同,可推断这两个小波包分解系数所包含的信号特征信息最少,因此,选取能量占比较大的其他 6 个小波包系数作为特征分量进行重构信号并提取特征。

2)声学信号的特征提取

当管道中状态发生改变时,在管道中传播的声学信号各频率段的能量发生了相应的变化,通过提取并进行分析此频率段能量的改变规律,实现管道堵塞状态的识别。

将提取多尺度空间能量特征的小波包分解的原理[15]是求解不同分解尺度上的信号能量,将求解所得能量值按照尺度顺序排列成特征向量,用以后续状态识别。

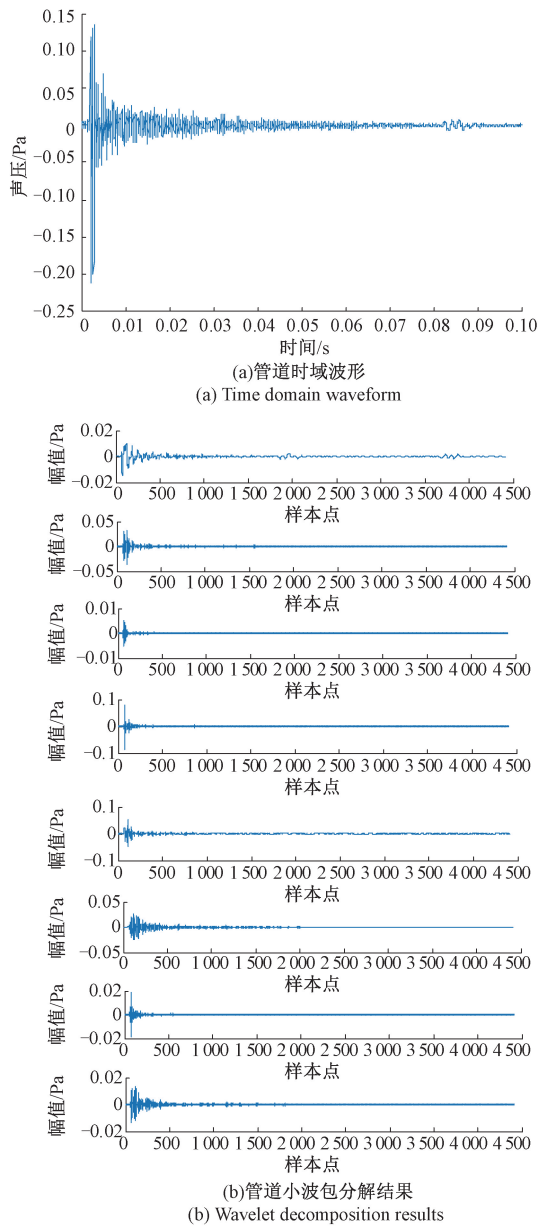


图 6　正常健康管道时域信号及其小波分解示意图

Fig. 6　Acoustic signal and its wavelet decompositions

在分解层次 j 上第 i 个节点的能量值 $E(i,j)$ 的计算式为:

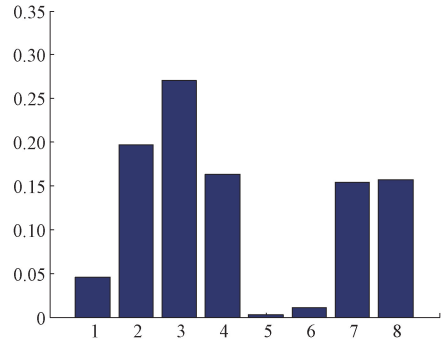
$$E(i,j) = \sum_{k\in Z} [p_s(n,j,k)]^2 \tag{12}$$

其中,Z 为采样点长度,$k=1,2,\cdots,Z$;$p_s(n,j,k)$ 是小波包变换系数,由 Parseval 能量积分公式[16]可知小波包变换系数的平方具有能量量纲。

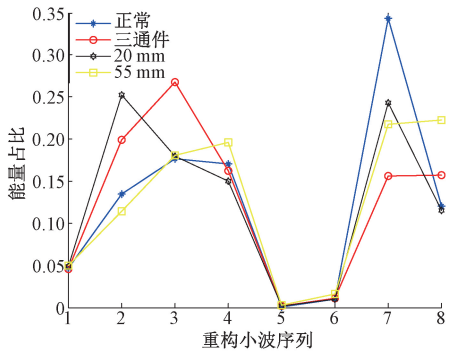
求分解层次 j 上第 i 个节点的能量值的概率 $p_i = E(i,j)/E$,定义小波包能量熵 P 为:

$$P = -\sum_{i=1}^{k} p_i \log p_i \tag{13}$$

近似熵用于量化时间序列波动的规律性,反映了时间序列中新信息发生的可能性,越复杂的时间序列对应



(a)正常管道声学信号能量分布
(a) Acoustic energy distribution in healthy pipe



(b)小波包分解系数能量
(b) Wavelet packet decomposition coefficient energy

图 7　小波分解能量分布对比

Fig. 7　Energy distribution of wavelet decompositions

的近似熵越大[15]。设存在 N 维连续时间序列,给定比较向量长度 m,相似度度量值 r,即 $U(N)=[u(1),u(i'),\cdots,u(N)]$, $i'=1,2,\cdots,N-m+1$。求出时间序列中矢量 $U(i')$ 与其余矢量 $U(N-i')$ 之间的最大距离,此最大距离与矢量个数 $N-m+1$ 的比值记为 $C_{i'}^{m}(r)$。先将 $C_{i'}^{m}(r)$ 取对数,再求其对所有 i' 的平均值,记为 $\Phi^{m}(r)$,即 $\Phi^{m}(r)=(N-m+1)^{-1}\sum_{i=1}^{N-m+1}\log(C_{i'}^{m}(r))$,再对 $m+1$,重复上述步骤得到 $\Phi^{m+1}(r)$。

由此,近似熵 $ApEn$ 定义式为:

$$ApEn=\Phi^{m}(r)-\Phi^{m+1}(r) \tag{14}$$

其中比较向量长度 m 一般取 1 或 2,相似度度量值 r 一般为 0.1~0.25 倍的序列标准差。

6 个小波包系数重构信号分别提取能量熵和近似熵,即形成 12 维的特征集合向量,部分结果如表 1 所示。

4.3　分类识别及结果分析

上述经过特征提取后正常健康管道(200 组)、含有三通件的健康管道(160 组)、20 mm 轻微堵塞状态管道(60 组)、55 mm 重度堵塞状态管道(40 组),选择各类样本的 2/3 样本作为训练集,剩余 1/3 样本作为测试集。

表 1　特征提取结果

Table 1　Feature extraction result

能量熵						
管道类别	1	2	3	4	7	8
正常管道	0.139	0.319	0.353	0.295	0.288	0.291
含三通件	0.154	0.329	0.300	0.292	0.367	0.110
轻微堵塞	0.146	0.347	0.308	0.284	0.343	0.249
重度堵塞	0.157	0.260	0.347	0.276	0.301	0.341
近似熵						
管道类别	1	2	3	4	7	8
正常管道	0.350	0.333	0.279	0.489	0.254	0.297
含三通件	0.324	0.381	0.274	0.438	0.183	0.233
轻微堵塞	0.266	0.276	0.148	0.348	0.139	0.365
重度堵塞	0.299	0.497	0.281	0.572	0.505	0.496

设定果蝇优化算法的初始参数,其中,果蝇种群总规模为 50,随机步长为 30,最大迭代次数为 200,C_1、C_2、C_3、C_4 和 g 的寻优范围为[0.01,1 000]。按照 Wang 等[17]定义的代价敏感分类器模型,样本数量少的惩罚因子取决于不平衡的比例。由领域知识所得,各类惩罚因子比例关系维持在 1.2∶1∶1∶1.3。

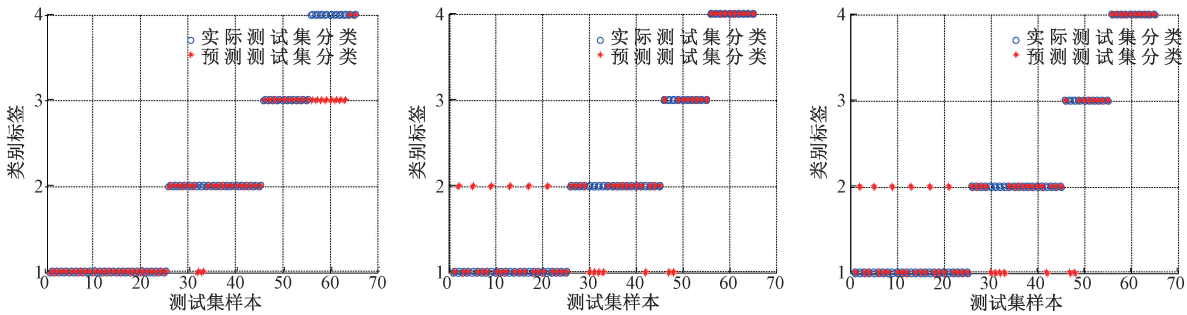
对特征集采用 FOA 优化的 CS-SVM 参数为 C_1 = 120.253 6,C_2 = 94.476 2,C_3 = 95.213 8,C_4 = 135.657 8,g = 110.873 2,总识别准确率为 83.636 4%;正常管道信号准确率 96.675 6%;含三通件管道信号准确率 90%;轻微堵塞管道信号准确率 87.5%;重度堵塞管道信号准确率 86.7%;含三通件管道信号错分为正常管道信号的错分率为 30.54%,重度堵塞管道信号错分为轻度堵塞管道信号的错分率为 10.76%。

对特征集未采用 FOA 优化的 CS-SVM 参数为 C_1 = 110.236 6,C_2 = 90.547 2,C_3 = 91.243 8,C_4 = 123.648 7,g = 100.845 3,总识别准确率为 79.636 5%;正常管道信号准确率 94.675 6%;含三通件管道信号准确率 86%;轻微堵塞管道信号准确率 87.5%;重度堵塞管道信号准确率 80.7%;含三通件管道信号错分为正常管道信号的错分率为 32.75%,重度堵塞管道信号错分为轻度堵塞管道信号的错分率为 20.46%。

采用相同的预处理方法和特征提取,经过 FOA 优化的传统 SVM 参数为 C = 53.281 6,g = 0.674 59,总识别准确率为 70.636 4%;正常管道信号准确率 90.654 3%;含三通件管道信号准确率 81.818 2%;轻微堵塞管道信号准确率 77.5%;重度堵塞管道信号准确率 30.879 6%;三通件管道信号错分为正常管道信号的错分率为 35.43%,重度堵塞管道信号错分为轻度堵塞管道信号的错分率为 60.42%。

不同方法在同一样本上的识别结果如图 8 所示,对于初始条件相同,即采用同样的预处理和特征提取,即相同特征集,本文所采用的 CS-SVM 方法与传统 SVM

相比，在含有三通件管道和 20 mm 管道识别准确率基本不变的情况下，提高了正常管道和 55 mm 管道识别准确率，在实际工况中，含有三通件管道信号极易错分为正常管道信号，55 mm 重度堵塞状态信号极易错分为 20 mm 轻微堵塞状态信号，所以本文方法是符合实际的。



(a) 传统SVM识别结果　(b) CS-SVM识别结果　(c) FOA-CS-SVM识别结果

(a) Traditional SVM recognition results　(b) CS-SVM recognition results　(c) FOA-CS-SVM recognition results

图 8　不同方法在同一样本集上的识别结果

Fig. 8　Identification results of different methods on the same sample set

以下验证所提的 CS-SVM 方法对不同比例的不同类样本识别的效果。实际工况中，排水管网运行状态多为正常运行状态，含三通件管道是为了改变流体方向，使主管道多出分支管，使用多少于正常直通管道，所以含三通件管道运行状态数据量少于管道正常运行状态。随着管道的服役年限增长，管道出现堵塞点后，随着水流面变窄，堵塞程度逐渐加深，形成重度堵塞，所以轻度堵塞管道运行状态数据样本量较重度堵塞管道运行状态数据样本数量多，但若重堵塞管道数据已多于全部数据的 1/2，说明管道已无需进行识别检测，已需要人工清堵。由此，分别采用不同类样本比为 1∶0.8∶0.2∶0.1，1∶0.8∶0.3∶0.2，1∶0.8∶0.4∶0.3，1∶0.8∶0.5∶0.4。对 CS-SVM 与传统 SVM 进行验证，声学信号采用相同的预处理方法，小波包分解重提取能量熵、近似熵，形成特征集。

由图 9 所示折线走向趋势来看，随着不同类样本不平衡比例的逐渐降低，3 种方法针对不同类别和总体样本的识别准确率都会随之升高，但提出的基于 FOA 的 CS-SVM 的方法相较于未优化的 CS-SVM 和基于传统 SVM 算法具有较高的识别准确率，且受样本不平衡性的影响较小，具有较强的泛化能力。

表 2~4 分别表示 FOA-CS-SVM 在不同比例样本条件下的识别结果、CS-SVM 在不同比例样本条件下的识别结果、传统 SVM 在不同比例样本条件下的识别结果，由对比观察所示，随着不同类样本不平衡比例的逐渐减小，3 种方法对于不同类样本的总体识别准确率随之增加，不同类的识别准确率基本随之增加，虽然在 20 mm 轻微堵塞管道的识别准确率有跳变；但针对总体而言，所提 FOA-CS-SVM 的方法相较于传统 SVM 算法和未优化的 CS-SVM 算法具有更高的识别准确率，对多类及类间不平衡样本的适应性较强。

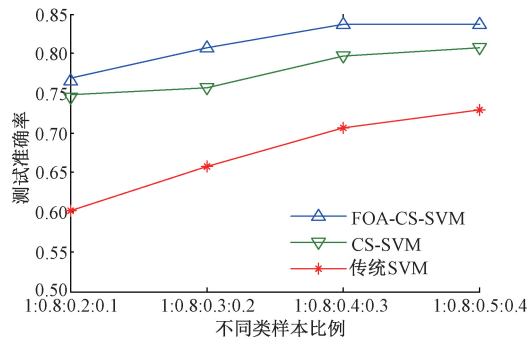


图 9　不同方法在不同比例样本集上的识别结果

Fig. 9　Identification results of different methods on different scale sample sets

表 2　FOA-CS-SVM 在不同比例样本条件下的识别结果

Table 2　The diagnosis results with FOA-CS-SVM in different proportions of imbalance sample

不同类样本比	识别准确率/%				
	正常	三通件	20 mm	55 mm	总体
1∶0.8∶0.2∶0.1	95.564 7	85.546 7	75.5	64.537 8	76.875
1∶0.8∶0.3∶0.2	95.275 6	87.818 2	76.459 1	78.259 8	80.679 2
1∶0.8∶0.4∶0.3	96.675 6	90.534 2	87.5	86.7	83.636 4
1∶0.8∶0.5∶0.4	97.818 2	90.818 2	75.672 1	87.2727	83.765 9

表 3　CS-SVM 在不同比例样本条件下的识别结果

Table 3　The diagnosis results with CS-SVM in different proportions of imbalance sample

不同类样本比	识别准确率/%				
	正常	三通件	20 mm	55 mm	总体
1∶0.8∶0.2∶0.1	92.564 7	80.654 3	74.438 1	60.897 3	74.875
1∶0.8∶0.3∶0.2	94.275 6	84.786 7	77.438 7	75.873 2	75.699 2
1∶0.8∶0.4∶0.3	94.675 6	86.765 2	87.5	80.743 5	79.636 5
1∶0.8∶0.5∶0.4	95.818 2	87.818 2	75.343 2	82.536 2	80.776 9

表 4　传统 SVM 在不同比例样本条件下的识别结果

Table 4　The diagnosis results with traditional SVM in different proportions of imbalance sample

不同类样本比	识别准确率/%				
	正常	三通件	20 mm	55 mm	总体
1∶0.8∶0.2∶0.1	90.387 6	76.657 8	70.765 6	20.456 3	60.272 7
1∶0.8∶0.3∶0.2	90.427 8	79.245 3	75.657 8	27.543 6	65.679 2
1∶0.8∶0.4∶0.3	90.654 3	81.818 2	77.532 4	30.879 6	70.636 4
1∶0.8∶0.5∶0.4	95.678 2	83.818 2	75.654 7	60.765 4	72.876 7

综上所述,本文 FOA-CS-SVM 方法与传统 SVM 算法和未优化的 CS-SVM 算法相比有效提高了对含有三通件管道和 55 mm 重度堵塞状态识别准确率,同时还能使正常管道和 20 mm 轻微堵塞状态识别准确率基本不变,使总体识别准确率提高。所提的基于 FOA 优化的 CS-SVM 模型针对类间样本不平衡问题可以有效的进行识别,比较适合处理故障类数据样本量较少以及不平衡数据样本集。

5　结　论

针对城市排水管道正常与堵塞故障状态在数据获取上的不平衡性造成的运行状态识别准确率下降的问题,根据排水管道内各运行状态下采集到的不平衡数据集,首先对不平衡数据集进行小波包分解,其次提取各个分解系数的能量熵、近似熵指标构建特征向量集合;采用 FOA 对不同类样本惩罚因子 C_m 和核函数参数 g 进行优化选取,即对 CS-SVM 模型优化,对排水管道的正常和堵塞状态识别,通过增大对少数类样本错分的惩罚代价,提升对少数类的识别准确率。通过一对一以二分类为基础的多分类识别方法,在决策层面解决不平衡的管道堵塞状态识别问题,在数据集样本数量较少的情况下,能有效处理不平衡程度对分类器的影响。由于准确精度与平衡数据时准确精度存在差距,在后续的研究中,可以探究是否可以应用深度学习或主动学习对不平衡数据集进行处理。

参考文献

[1] 李若晗. 城市污水管道检测、评价与影响因素研究[D]. 北京:清华大学,2016.
LI R H. Research on the detection, evaluation and influencing factors of urban sewage pipelines [D]. Beijing: Tsinghua University,2016.

[2] 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):110-118.
LANG X M, LI P, CAO J T, et al. Pipeline leak detection and location based on unbalanced data processing[J]. Journal of Hunan University (Natural Science),2018, 45(2):110-118.

[3] 闫慈,田翔华,阿拉依阿汗,等. 基于重采样技术在医学不平衡数据分类中的应用研究[J]. 中国卫生统计,2018,35(2):177-180,185.
YAN C, TIAN X H, ALAYI A H, et al. Application of resampling technique in medical imbalance data classification[J]. Chinese Journal of Health Statistics, 2018, 35 (2): 177-180,185.

[4] 许玉格,孙称立,赖春伶,等. 基于不平衡学习的集成极限学习机污水处理故障诊断[J]. 化工学报,2018, 69(7):3114-3124.
XU Y G, SUN CH L, LAI CH L, et al. Fault diagnosis of wastewater treatment based on integrated learning machine based on unbalanced learning[J]. Journal of Chemical Industry and Engineering, 2018, 69 (7): 3114-3124.

[5] 谭洁帆,朱焱,陈同孝,等. 基于卷积神经网络和代价敏感的不平衡图像分类方法[J]. 计算机应用,2018, 38(7):1862-1865,1871.
TAN J F, ZHU Y, CHEN T X, et al. Unbalanced image classification based on convolutional neural network and cost sensitivity[J]. Journal of Computer Applications, 2018, 38 (7): 1862-1865, 1871.

[6] 文玉梅,张雪园,文静,等. 依据声信号频率分布和复杂度的供水管道泄漏辨识[J]. 仪器仪表学报,2014, 35(6):1223-1229.
WEN Y M, ZHANG X Y, WEN J, et al. Leakage identification of water supply pipeline based on frequency distribution and complexity of acoustic signal [J]. Chinese Journal of Scientific Instrument, 2014, 35 (6): 1223-1229.

[7] 康守强,许林虎,王玉静,等. 基于 FOA-MKSVM 的滚动轴承故障分类方法[J]. 仪器仪表学报,2015, 36(5):1186-1192.
KANG SH Q, XU L H, WANG Y J, et al. Classification method of rolling bearing fault based on FOA-MKSVM[J]. Chinese Journal of Scientific Instrument, 2015, 36(5): 1186-1192.

[8] 向阳辉,张干清,庞佑霞,等. 多分类 SVM 的代价敏感加权故障诊断方法[J]. 振动、测试与诊断,2015, 35(6): 1116-1122,1202-1203.
XIANG Y H, ZHANG G Q, PANG Y X, et al. Cost-sensitive weighted fault diagnosis method for multi-class SVM [J]. Journal of Vibration, Measurement and Diagnosis, 2015, 35(6): 1116-1122,1202-1203.

[9] 刘永斌,何清波,孔凡让,等. 基于 PCA 和 SVM 的内燃机故障诊断[J]. 振动、测试与诊断,2012,32(2): 250-255.

LIU Y B, HE Q B, KONG F R, et al. Fault diagnosis of internal combustion engine using PCA and SVM[J]. Journal of Vibration, Measurement and Diagnosis, 2012, 32(2): 250-255.

[10] 李艳霞,柴毅,胡友强,等. 不平衡数据分类方法综述[J]. 控制与决策,2019,34(4):673-688.
LI Y X, CHAI Y, HU Y Q, et al. Overview of unbalanced data classification methods[J]. Control and Decision, 2019, 34(4): 673-688.

[11] 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):108-114.
HE D W, PENG J B, HU J H, et al. Fault diagnosis of CS-SVM bearing based on improved FOA optimization [J]. Journal of Vibration and Shock, 2018, 37(18): 108-114.

[12] ZIANI R, FELKAOUI A, ZEGADI R. Bearing fault diagnosis using multiclass support vector machines with binary particle swarm optimization and regularized Fisher's criterion [J]. Journal of Intelligent Manufacturing, 2017, 28(2): 405-417.

[13] PAN W T. A new fruit fly optimization algorithm: Taking the financial distress model as an example [J]. Knowledge-Based Systems, 2012, 26(2): 69-74.

[14] 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 pipeline jamming fault based on generalized Fisher-mutual information [J]. Journal of Electronic Measurement and Instrumentation, 2018, 32(11): 1-8.

[15] 沙洲,杨洋,刘颖华,等. 小波信息熵在输水管道泄漏检测技术中的应用[J]. 电子测量与仪器学报,2018, 32(7):151-156.
SHA ZH, YANG Y, LIU Y H, et al. Application of wavelet information entropy in water pipeline leak detection technology [J]. Journal of Electronic Measurement and Instrumentation, 2018, 32(7): 151-156.

[16] 马朝永,盛志鹏,胥永刚,等. 基于自适应频率切片小波变换的滚动轴承故障诊断[J]. 农业工程学报, 2019(10):34-41.
MA CH Y, SHENG ZH P, XU Y G, et al. Fault diagnosis of rolling bearings based on adaptive frequency slice wavelet transform[J]. Transactions of the Chinese Society of Agricultural Engineering, 2019(10): 34-41.

[17] WANG H, WANG N, YEUNG D Y. Collaborative deep learning for recommender systems [C]. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2014: 1235-1244.

作者简介

王菲,2016 年于石家庄学院获得学士学位,现为昆明理工大学硕士研究生,主要研究方向为机器学习研究。
E-mail:2279677714@qq.com

Wang Fei received B. Sc. from Shijiazhuang University in 2016. Now she is a M. Sc. candidate at Kunming University of Science and Technology. Her main research direction is machine learning research.

冯早(通信作者),2009 年于英国纽卡斯尔大学获硕士学位,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伤检测技术及应用研究、数据挖掘、机器学习算法研究。
E-mail:6483975@qq.com

Feng Zao(Corresponding author) received M. Sc. from the 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 the research of non-destructive testing technology, application research, data mining and machine learning algorithm based on acoustics.