基于灰关联信息熵的 SVM 决策树轴承故障分类*

陈　伟，　贾庆轩，　孙汉旭
（北京邮电大学自动化学院　北京，100876）

摘要　为提高轴承故障诊断的准确率，以灰色关联理论和信息熵理论为基础，提出了基于灰关联信息熵提取属性特征的支持向量机决策树多故障分类器。该分类器可以实现对轴承的多故障类型的分类，并对轴承的各类故障进行了分类实验。验证结果表明，该方法可有效地进行故障状态识别，达到了准确进行机械系统多故障诊断的目的。

关键词　支持向量机(SVM)；决策树；轴承；灰色关联；信息熵
中图分类号　TP391；TH133.3

引　言

在机械系统运行时，其部件及零件的故障 30% 是由轴承故障引起的，若能较快地发现并预测轴承故障和类型，及时采取相应的措施，将故障消除在萌芽状态，对设备完善、人员安全都将有很强的实际意义。近年来各种故障识别研究方法中，支持向量机(SVM) 是一种仅用少样本情况下，对机器学习问题所开展的一种先进的智能故障识别理论体系，该系统的特点是能够针对小样本数据、高维及非线性等实际问题[1]进行故障识别。正因为其所具有的这些优势，支持向量机目前正在成为设备智能故障识别理论体系的新方向。最初支持向量机是为了解决二类分类问题而建立的模型，为了使其能够解决多分类问题，已经提出了多种处理思路，主要有：一对一分类、一对多分类、有向非循环图支持向量机、决策树支持向量机等。虽然支持向量机决策树方法从训练速度还是分类精度都优于其他方法，但是也存在故障辨识率不高等问题[1]，尤其无法满足实际中遇到的多类故障问题。

为了进一步提高 SVM 决策树故障分辨率，笔者采用灰色关联理论融合信息熵理论的方法对轴承故障样本进行特征属性信息提取，并采用改进的一对一决策树结构，构造支持向量机多分类器，以改善在轴承多故障辨识中的分类精度不高的现状，提高其故障辨识率。最后，通过轴承故障模拟实验对该分类器的分类精度进行了验证，证实了该研究方法的有效性。

1　信息熵理论

目前，基于信息熵理论的特征属性提取算法是国际上最有影响的分类学习方法之一。Claude E Shannon 把信息(熵)定义为离散随机事件的出现概率，表示信息的不确定性，可以作为对信息价值的判断标准[2]。信息熵分类算法采用分枝策略，在决策过程中，以所提取的信息特征的增益大小作为其属性选择的启发式函数，该算法如下。

假定 S 是 s 个样本数据的集合，设定有 m 个不同的类别 $C_i(i=1,2,\cdots,m)$，s_i 为类别 C_i 的样本数，对于 C_i 这样一个样本分类，最终的期望信息为 $I(s_1,s_2,\cdots,s_m)=-\sum_{i=1}^{m}(p_i\log p_i)$，也称样本熵，其中 p_i 为样本 S 属于 C_i 的概率，一般可以用 s_i/s 来估计。

若设定具有 n 个不同的值$\{a_1,a_2,\cdots,a_n\}$，其属性为 A，则可用 A 将 S 进行划分，划分成 n 个子集$\{s_1,s_2,\cdots,s_n\}$，其中 S_i 包含 S 中在属性 A 上值为 a_i 的样本。若设定 A 为最好的分裂属性作为测试属性，则 n 个子集对应于集合 S 的节点所生长出来的分支。

设 s_{ij} 是子集 s_j 中属于样本分类 C_i 的样本熵数，则根据 A 划分的子集的熵

$$E(A)=-\sum_{j=1}^{n}w_j\cdot I(s_{1j},s_{2j},\cdots,s_{mj}) \qquad (1)$$

* 国家自然科学基金资助项目(61175080)
收稿日期：2012-12-26；修改稿收到日期：2013-02-26

其中：$w_j=s_{1j}+s_{2j}+\cdots+s_{mj}/s$ 为第 j 个子集的权；$I(s_{1j},s_{2j},\cdots,s_{mj})=-\sum_{i=1}^{m}(p_{ij}\log_2 p_{ij})$ 为给定子集 s_j 的期望值，$p_{ij}=s_{ij}/s_j$ 为 s_j 中的样本，属于样本分类 C_i 的概率熵越小的子集所划分的类别纯度越高。

通过所期望的信息及熵值求得所对应的信息增益，则在 A 上分支将能得到的信息增益为

$$G(A)=I(s_1,s_2,\cdots,s_m)-E(A) \tag{2}$$

使用式(2)就可以计算出每个子集 s 的属性 $G(A)$ 的增益，然后选取其中最高的信息增益的属性作为给定集合 S 的测试属性，对被选取的测试属性创建一个结点，并以该属性为标记，对该属性的每个属性值创建一个分支，并以此来划分样本。

基于信息熵的分类算法虽然有效，但该算法有一定的局限性[3]。应用中在分类决策时通常偏向选择取值较多的属性，而实际中取值较多的属性往往并不是最重要的，即按照信息增益最大的原则，被分类算法列为应选取的属性有时对其进行测试不会提供太多的信息，因此需要将其进行改进再加以应用。

2　灰色关联理论改进信息熵分类算法

如果要改进信息熵分类算法，首要考虑的就是优化对属性的选择标准，可以通过对信息熵的公式加权来加强重要属性的标注，降低非重要属性的标注。灰色关联理论是通过分析数据列因素之间的相似或者相异程度来衡量数据行接近的程度，其分析过程和结果反映了不同特征在分类决策中的重要性[4]，因此，通过灰色关联分析来改进信息熵分类算法是可行的。

2.1　灰色关联度的计算

采用相对灰色关联度来标定属性的权值，将滚动轴承的运行过程作为一个灰色系统，进行其关联度分析时，所采取的基本思路是依照灰色关联度的大小，来加强其重要属性的标注，同时还要降低非重要属性的标注，轴承运行状态的灰色关联度计算步骤如下。

1) 首先针对正常状态和故障状态的轴承的所呈现的振动信号中提取其特征参数 X_i，建立其特征向量矩阵作为标准故障模式 B_i 为

$$B_i=\begin{bmatrix}B_1\\B_2\\\vdots\\B_m\end{bmatrix}=\begin{bmatrix}X_1(1)&X_1(2)&\cdots&X_1(n)\\X_2(1)&X_2(2)&\cdots&X_2(n)\\\vdots&\vdots&\cdots&\vdots\\X_m(1)&X_m(2)&\cdots&X_m(n)\end{bmatrix} \tag{3}$$

2) 将测取的待检轴承的故障特征参数 X_T 构成待检模式 D_T 为

$$D_T=X_T=[X_T(1),X_T(2),\cdots,X_T(n)] \tag{4}$$

3) 计算待检模式与标准模式中各类故障类型的关联度

定义第 j 个特征的绝对差为 $\Delta_{ij}=|X_T(j)-X_i(j)|$，则可得待检特征数列 $X_i(j)$ 与故障特征数列 $X_T(j)$ 的关联系数为

$$\xi_i(j)=\frac{\min\limits_i\min\limits_j\Delta_{ij}+\rho\max\limits_i\max\limits_j\Delta_{ij}}{\Delta_{ij}+\rho\max\limits_i\max\limits_j\Delta_{ij}} \tag{5}$$

其中：ρ 的取值范围为 $[0,1]$，通常取 $\rho=0.5$；$\min\limits_i\min\limits_j\Delta_{ij}$ 为两级最小值，$\max\limits_i\max\limits_j\Delta_{ij}$ 为两级最大差。

由于关联系数结果较多，不便于比较，因此通常采用关联度这个概念，记为 $r(X_0,X_i)$。在计算目标灰色关联度时，各特征的重要性不同，应该分配不同的权重 $a(j)$，可定义加权关联度为

$$R_i=\sum_{j=1}^{N}\xi_i(j)a(j)\quad(i=1,2,\cdots,M)$$

其中

$$a(j)=\xi_i(j)\Big/\sum_{j=1}^{N}\xi_i(j) \tag{6}$$

由计算过程可见，灰色关联系数体现了待研究序列的每个属性与样本的相关程度，而灰色关联度则从整体上反映了研究序列与样本的相关性。

2.2　信息熵分类算法的改进

利用灰色关联度对属性信息熵的计算公式加权，以加强重要属性的标注，将式(1)(2)改进为

$$E(A)=-\sum_{j=1}^{n}w_jR_iI(s_{1j},s_{2j},\cdots,s_{mj}) \tag{7}$$

$$G(A)=I(s_1,s_2,\cdots,s_m)-E(A)=$$

$$\sum_{j=1}^{n}w_jR_i\sum_{i=1}^{m}(p_{ij}\log_2 p_{ij})-\sum_{i=1}^{n}(p_i\log_2 p_i) \tag{8}$$

其中：R_{i0} 为各子属性与根属性之间的相对灰色关联度。

改进后的分类准则可以在分类器训练时有效地提取样本中的重要属性信息，在分类时以样本重要属性作为优先决策因素，以此提高故障辨识精度。

3　支持向量机多分类器模型建立

构建多分类故障分类器，通常多采用一对一方法、一对多方法。

1) 一对一方法在所有 N 类之间构造所有可能的两类分类器，通常采用投票方法来确定分类结果，此时需要构造 $N(N-1)/2$ 个两类分类器。一对一

方法分类方法简单，效果比较好，但对于分类类型较多时，计算量增大导致代价太大。

2) 一对多方法构造 N 个两类分类器，通过比较 N 个分类器的输出值来判定分类结果，该方法计算代价较小，但是其分类效果一般[5-6]。

鉴于以上方法的缺陷，笔者拟在“一对一”多分类算法的基础上，结合上述提出的灰色关联理论改进信息熵理论的决策算法，构造一种既能够提高已知少量样本的重复利用度，又能显著提高分类精度的多分类器。

建立分类器之前首先对样本属性进行灰关联信息熵处理：样本包括轴承的 4 类状态，正常状态 X_1、内圈故障 X_2、外圈故障 X_3、滚动体故障 X_4；以 X_1 样本为参考标准，X_2，X_3，X_4 的每项特征属性代入式(5)与样本集 X_1 进行灰关联分析；使用式(8)求得其每项特征属性的灰关联信息增益，对特征属性进行数值到信息熵的转换。然后，假设要区分 n 类故障：a. 根据故障类型的平均分配原则，把样本按照不同的故障类型先分为两组，训练一个二类分类器；b. 把故障类型的在进行分组变换后，重新训练一个新的二类分类器，依次循环，可知第一级共可以训练 $\frac{1}{2}C_n^2$ 个分类器：SVM_1，SVM_2，…；c. 在每一个 SVM 中，继续以平均分配原则的方式训练下一级的 SVM，当出现单个故障类型时停止训练(如果所训练的故障类型比较多时，在第 1 级分组中，可以不必每种分组情况都训练，只训练 3 或 5 个，这样可以减少 SVM 的数量，从而节省了训练和分类的时间)。笔者以 4 分类器为例构造起一个分类器，构造过程如图 1 所示。首先，在第 1 级将通过式(7)求得的四类样本熵分成 X_1，X_2 和 X_3，X_4 两大类训练 SVM_1；然后，在第 2 级用样本熵 X_1 和 X_2 训练 SVM_{11}，用样本熵 X_3 和 X_4 训练 SVM_{12}，完成第 1 部分分类器的训练。改变样本排列顺序，用相同的方法完成第 2 部分和第 3 部分分类器的训练。

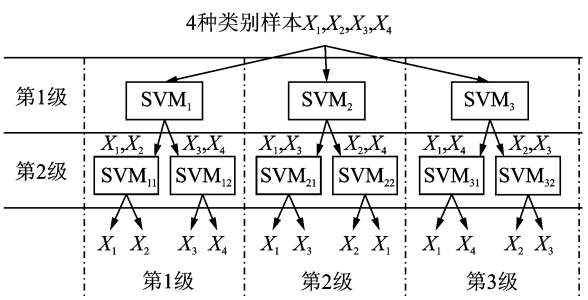

图 1 支持向量机故障诊断四分类器

训练完毕完成了支持向量机决策树的建立。进入故障分类测试阶段：当有一个待分类数据 X_i，仍然需要首先代入式(5)与样本集进行灰关联分析，使用式(8)求得特征属性值对应的灰关联信息熵。然后，输入到上面训练好的多分类 SVM 中，数据会同时在每一部分的分类器中从上向下运算，并得到一个分类结果。

由于训练样本数据、待分类数据本身的好坏，以及分类累积误差的存在，每部分分类器得出的分类结果可能不尽相同。在确定故障类型时，分类器训练时有效地选择重要属性信息，可以提高故障辨识精度。

4 实 验

本实验所搭建的轴承故障模拟试验台，是由机械驱动装置、加载机构、固定装置和机座 4 部分组成。电机型号为 Y100L2-4；额定功率为 3 kW；额定转速为 1 420 r/min。轴承型号为 2612，该轴承在实验时其外圈是固定不动的。利用加速度传感器拾取 4 种情况(正常轴承、轴承外圈故障及轴承内圈故障、轴承滚动体故障)下的轴承振动信号。采样频率为 f_s=2 kHz。振动信号采集装置包括加速度传感器、藕合器、数据采集分析仪等。振动传感器的响应频率为 22 kHz。测量滚动轴承所在主轴的实际转速所采用的传感器是电涡流位移传感器，实验台如图 2 所示。

人为在轴承内外圈及滚动体上加工出 0.15，0.30，0.45 mm3 种故障孔，用以模拟轴承的局部故障信号，轴承运行状态如表 1 所示。

由于轴承振动信号的各种时、频域特征参量对故障信息的表达能力各有侧重，只有进行综合分析才能得到比较完整的故障信息。诊断模型选用 4 个时域特征参量：峰值指标、峭度指标、裕度指标、波形指标；3 个频域特征参量：功率谱重心、功率谱方差、

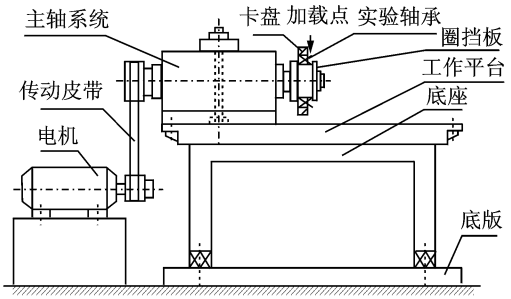

图 2 滚动轴承故障模拟实验台示意图

表 1　轴承的运行状态

数据集	运行状态	训练样本个数	测试样本个数	故障点直径/mm	转速/(r·min^{-1})
1	正常	10	20	0.15	
	内环故障	10	20		
	外环故障	10	20		
	滚动体故障	10	20		
2	正常	10	20	0.3	1 420
	内环故障	10	20		
	外环故障	10	20		
	滚动体故障	10	20		
3	正常	10	20	0.45	
	内环故障	10	20		
	外环故障	10	20		
	滚动体故障	10	20		

谐波因子。由于这些特征参量存在着量纲差别，因此首先进行归一化处理，然后使用式(8)进行特征参数的灰关联信息熵计算，部分计算后的数据如表 2 所示。

在非线性条件下，支持向量机分类问题的最优分类函数为

$$f(x)=\mathrm{sgn}\Big[\sum_{i=1}^{n}a'_i\,y_iP(x_i,x)+b\Big] \tag{9}$$

其中：$P(x,y)$称为核函数。

目前所用的核函数有：线性核函数、多项式核函数、径向基核函数和 Sigmoid 函数等。其中径向基函数由于其待定参数少，并且对应的特征空间可以是无穷维的，有限的数据样本均可取得良好的分类效果，因而使用最为广泛[7-8]。

表 2　轴承测试特征参数灰关联信息熵部分数据

故障模式	采样频率/kHz	峰值指标	峭度指标	裕度指标	波形指标	功率谱重心	功率谱方差	谐波因子
正常(X_1)	2	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.000 0	0.012 1
内圈故障(X_2)	2	0.180 2	0.208 9	0.283 1	0.092 4	0.873 4	0.137 2	0.082 1
		1.000 0	0.406 9	1.000 0	0.423 8	0.869 2	0.157 8	0.512 9
外圈故障(X_3)	2	0.472 7	0.287 7	0.408 2	0.178 8	0.646 7	0.114 6	1.00 00
		0.888 5	1.000 0	0.871 7	0.559 6	0.830 2	1.000 0	0.158 3
滚动体故障(X_4)	2	0.584 1	0. 792 1	0.640 6	1.000 0	1.000 0	0.579 8	0.000 0
		0.104 7	0.174 1	0.075 3	0.064 1	0.711 3	0.565 1	0.091 3

选择使用径向基核函数计算支持向量内积，径向基函数形式为

$$K(x,y)=\exp\left[-\frac{\|x-y\|^2}{2\sigma^2}\right] \tag{10}$$

确定误差惩罚参数 C 和核参数 σ。经过分析比较，将误差惩罚参数 C 选为 100，核参数 σ 选为 0.1 时，此时多故障分类器的推广能力最好。

将所采集的正常轴承 20 组，滚动体故障 20 组，外圈故障 20 组，内圈故障 20 组共 80 组特征向量，选其中 60 组作为训练样本，另 20 组作为验证样本。带入灰色关联理论与决策树理论相融合的四分类的多分类器支持向量机中，训练完毕后进入测试阶段，把剩下的 120 组测试数据分为 6 个数据集，做 6 次测试，每次测试 20 组数据，训练样本数为 40。分类结果见表 3 所示。

表 3　基于 SVM 的滚动轴承多故障分类结果表

数据集	测试样本数	正确分类数	正确率 %
1	20	20	
2	20	19	
3	20	18	95.8±0.82
4	20	20	
5	20	19	
6	20	19	

利用数理统计的知识可以计算出，此多分类器构造的支持向量机针对本研究课题的分类正确率达到(95.8±0.82)%。由此证明，笔者所建立的轴承多故障分类器具有良好的分类能力和分类精度。

5　结束语

构建了支持向量机与信息熵理论、灰色关联理论相融合的故障诊断识别技术。通过实验证实了所建立的根据灰色关联度的大小加强决策树重要属性的标注，降低非重要属性的标注的 SVM 多故障分类器，算法简单，分类效果好，且只需要较少的样本就能达到分类的要求。由于重复有效利用了有限样本，提高了少样本的重复利用率，近似于增加了样本的数量，增加了识别的精确性。该方法不但可以识别诊断滚动轴承故障，还可以适用其他机械设备的故障类型的诊断与识别中。

参　考　文　献

[1]　Viadimir N V. 统计学习理论的本质[M]. 张学工，译. 北京：清华大学出版社，2012：163-168.

[2] 赵志宏,杨绍普.基于小波包变换和样本熵的滚动轴承故障诊断[J]. 振动、测试与诊断,2012,32(4):640-644.
Zhao Zhihong, Yang Shaopu. Roller bearing fault diagnosis based on wavelet packet transform and sample entropy[J]. Journal of Vibration, Measurement & Diagnosis, 2012,32(4):640-644. (in Chinese)

[3] 苏文胜,王奉涛,朱泓,等.基于小波包样本熵的滚动轴承故障特征提取[J]. 振动、测试与诊断,2011,31(2):162-166.
Su Wensheng, Wang Fengtao, Zhu Hong, et al. Feature extraction of rolling element bearing fault using wavelet packet sample entropy[J]. Journal of Vibration, Measurement & Diagnosis, 2011,31(2):162-166. (in Chinese)

[4] 刘胧,刘虎沉,林清恋.基于模糊证据推理和灰色关联理论的 FMEA 方法[J]. 模糊系统与数学,2011,25(2):71-80.
Liu Long, Liu Huchen, Lin Qinglian. An improved FMEA using fuzzy evidential reasoning approach and grey theory[J]. Fuzzy Systems and Mathematics, 2011,25(2):71-80. (in Chinese)

[5] 袁胜发,褚福磊.支持向量机及其在机械故障诊断中的应用[J].振动与冲击, 2007, 26 (11): 29-34.
Yuan Shengfa, Chu Fulei. Support vector machines and its applications in machine fault dlagnosis[J]. Journal of Vibration and Shock, 2007, 26 (11):29-34. (in Chinese)

[6] 何学文,卜英勇.基于小波包分解和支持向量机的机械故障诊断方法[J].机械强度,2004,26(1):20-24.
He Xuewen, Bu Yingyong. Method of machinery fault dlagnosis based on wavelet packet decomposition and support vector machine[J]. Journal of Mechanical Strength,2004,26(1):20-24. (in Chinese)

[7] 孙昌儿,刘秉瀚.一种新的 SVM 决策树[J].福州大学学报:自然科学版,2007,35(3):361-364.
Sun Changer, Liu Binghan. A new SVM decision tree[J]. Journal of Fuzhou University: Natural Science Edition, 2007,35(3):361-364. (in Chinese)

[8] 王计生,喻俊馨,黄惟公.小波包分析和支持向量机在刀具故障诊断中的应用[J].振动、测试与诊断,2008,28(3):273-276.
Wang Jisheng, Yu Junxin, Huang Weigong. Application of wavelet package analysis and support vector machine to fault diagnosis of cutting tool[J]. Journal of Vibration, Measurement & Diagnosis, 2008, 28(3):273-276. (in Chinese)

第一作者简介:陈伟,男,1981 年 1 月生,博士研究生。主要研究方向为物流机械设备运行监测与振动故障处理。曾发表《邮政及物流设备设计》(北京:人民邮电出版社,2011 年)等论著。
E-mail:cschwbeijing@126.com