

基于改进分类器动态选择算法的滚珠丝杠副状态识别^①

文 娟^②

(浙江工业大学机械工程学院 杭州 310023)

(恒丰泰精密机械股份有限公司 温州 325000)

摘 要 为提升滚珠丝杠副的性能状态识别精度,提出一种改进的分类器动态选择算法。该算法借助邻域成分分析(NCA),准确并自适应地定义测试样本的邻域,无需选择距离度量方式,从而更加准确地衡量多分类器系统中各子分类器对于测试样本进行正确分类的潜力,解决了传统分类器动态选择算法精度受限于距离度量方式选择是否合适的问题。将所提出的分类器动态选择算法应用于滚珠丝杠副状态识别中,首先利用 AdaBoost 算法离线训练反向传播(BP)神经网络集合,然后依据实时信号特征,采用改进的分类器动态选择算法从分类器集合中选取最合适的子分类器进行状态鉴定,从而实现更好的识别效果。实验结果表明,提出方法的状态识别准确率能够达到 97.22%,高于 BP 神经网络、AdaBoost 与传统分类器动态选择算法,且对于不同的性能状态均有较高的识别精度。

关键词 分类器动态选择;邻域成分分析(NCA);状态识别;滚珠丝杠副;多分类器系统

滚珠丝杠副是一种能够将旋转运动转换为直线运动的装置,具有高精度和高效率的特性,因而被广泛应用于数控机床的进给系统中^[1]。作为机床的关键部件,滚珠丝杠副的性能状态直接影响着整个数控机床的可靠性与安全性。因此,对其进行在线监测,通过实时检测数据判定滚珠丝杠副的退化状态,进而制定合理的维护维修计划,能够确保产品的加工质量,保证生产过程的安全运作,并降低维修成本,具有重要意义^[2-3]。

现有的滚珠丝杠副状态监测方法通常可以分为信号处理方法和智能模型方法。前者通过信号处理技术分析监测信号,从而判定滚珠丝杠副的性能状态。Tsai 等人^[1]研究发现,由于预紧力的摩擦作用,含预紧力滚珠丝杠副的滚珠经过频率小于无预紧力滚珠丝杠副的经过频率,因而提出一种通过监测滚珠经过频率监测预紧力丧失的方法。Chang 等人^[4]结合经验模态分解和多尺度熵对滚珠丝杠副驱动电

机的力矩信号进行处理,得到能够反映滚珠丝杠副预紧力水平的信号特征量。与这类方法不同,智能模型方法无须任何先验知识,通过机器学习算法建立监测数据与滚珠丝杠副退化状态的关系,从而自动识别滚珠丝杠副的性能状态,并逐渐成为主流的状态监测方法。Jin^[5]采用体制矩阵限制玻尔兹曼机提取滚珠丝杠副的退化特征,然后利用自组织映射评估其健康状态。Li 等人^[6]采用支持向量机和高斯过程回归建立滚珠丝杠副的状态识别模型。吴希曦等人^[7]将小波包提取的特征输入超球面支持向量机中,实现了丝杠的状态识别。Huang 与 Hsieh^[8]利用内置信号表征滚珠丝杠副的健康状态,并采用支持向量机作为分类器实现状态识别。

智能模型状态监测方法中,机器学习算法,也就是分类器的性能在很大程度上影响着状态识别的准确度^[9-10]。相对于单个分类器而言,多分类器系统能够克服单个分类器的缺点,并提高分类准确率^[11-13]。

① 国家自然科学基金(51475425),绍兴市“揭榜挂帅”制科技项目(2021B41006)和浙江省博士后科研项目择优(273426)资助项目。

② 女,1990 年生,博士;研究方向:机械故障诊断与寿命预测;联系人,E-mail: juanwen@zjut.edu.cn。

(收稿日期:2022-11-07)

通常,构建多分类器系统的方式有 2 种:分类器融合与分类器动态选择。生成由多个子分类器构成的分类器集合后,前者通过融合每一个子分类器的输出得到测试样本的标签,但分类器集合中通常存在一些不准确的分类器,这些分类器会影响最终输出的准确率。与分类器融合方式不同,分类器动态选择方法在对每一个测试样本进行分类时,不再考虑所有子分类器的输出结果,而是根据测试样本的特点从前述分类器集合中选取一个最合适的子分类器对其进行分类。因为不同的测试样本通常会涉及到不同的分类难点,因此采用不同的分类器对其进行分类比采用分类器融合方法确定其标签更为合理。

目前,有多种基于不同准则的分类器动态选择算法,如局部类精度(local class accuracy, LCA)^[14]、K 近邻上限法^[15]以及数据处理组选择法^[16]等。虽然无法证明哪一种分类器动态选择方法对于所有的分类问题效果都优于其他方法,但是通常采用简单的 LCA 方法就能获得与复杂方法相似甚至更好的分类性能^[17]。LCA 算法通过计算测试样本与训练样本或验证样本的距离,将与其距离最近的 K 个近邻定义为邻域,计算每个子分类器对于其邻域的分类准确率,选取准确率最高的子分类器对该测试样本进行分类,从而提高多分类器系统的分类性能。

然而,在很多情况下, LCA 算法也无法从分类器集合中找出最适合测试样本的子分类器。这是由于使用 K 近邻准则进行邻域定义的限制性,其邻域定义的准确与否依赖于测试样本与训练样本之间距离衡量的准确性。因此,如何选择合适的距离度量方式将会影响邻域的范围,进而影响 LCA 的性能。在作者之前的研究中,基于滚珠丝杠副的状态监测数据,采用分类器动态选择的方式构建了多分类器系统,即通过 LCA 算法针对每个未知状态样本,从分类器集合中选取最合适的子分类器进行滚珠丝杠副状态识别。结果表明,通过 LCA 算法能够较为准确地识别正常和失效 2 种状态,但对于退化状态的判定效果不是很理想,受限于 LCA 算法的性能。

为解决这一问题,本文提出一种新的分类器动态选择方法:邻域局部类精度(neighborhood local class accuracy, NLCA)。该方法通过引入一种名为

邻域成分分析(neighborhood components analysis, NCA)^[18]的度量学习方法来定义测试样本的邻域,以提升邻域范围定义的准确度。NCA 通过度量学习,将原始数据映射到另一个空间,在这个空间内 K 近邻分类器(k-nearest neighbor, KNN)具有很高的精度,即针对每个测试样本,使用 KNN 找到其真实近邻的概率非常高。因此,采用 NCA 能够准确并且自适应地确定测试样本的邻域,无需选择距离度量方式,进而提高 LCA 算法的精度,最终达到提高分类精度的目的。最后,对所提 NLCA 方法在机器学习标准数据集和滚珠丝杠副数据集上进行了验证,说明了该方法的有效性与优越性。

1 局部类精度算法

LCA 是一种基于局部分类准确率的分类器动态选择方法。当对未知标签的样本(测试样本)进行分类时, LCA 算法会计算分类器集合中每一个子分类器对于与该样本相邻的局部区域所包含样本的分类性能,从而选取准确率最高的子分类器来预测该样本的标签。最常用的邻域定义方法是通过计算测试样本与验证集或训练集的距离寻找该测试样本的 K 个最近邻作为其邻域。在此基础上, LCA 通过计算各子分类器对于邻域中一些特定样本的分类准确率来衡量其正确输出测试样本标签的潜力,具体的实现步骤如下所述。

(1) 输入标签种类 Ω 、子分类器集合 C 、训练集 X 和测试集 T 、邻域尺寸 K 。

(2) 对于每个测试样本 t , 进行以下操作:

使用所有子分类器对 t 进行分类, 如果所有子分类器的输出标签相同, 则返回该标签作为样本 t 的类别; 反之, 则针对每个子分类器 $c_m \in C$, 采用以下步骤计算其竞争力, 并选取最适合 t 的子分类器。

1) 采用子分类器 c_m 预测样本 t 的类别, 得到其标签 $\omega_j = c_m(t)$;

2) 从训练集 X 中找到样本 t 的 K 个最近邻, 采用 c_m 分别对这 K 个最近邻进行分类, 得到其标签。从这 K 个最近邻中选择标签与 ω_j 相同的训练样本, 使用 ψ 代表这些样本所定义的域;

3) 计算子分类器 c_m 在 ψ 上的分类正确率 $LCA(m, j)$;

4) 采用下式选择最适合测试样本 t 的分类器 c_t , 即选择在 ψ 上识别精度最高的分类器:

$$c_t = \operatorname{argmax}_m \{LCA(m, j)\} \quad (1)$$

5) 使用 c_t 对测试样本 t 进行分类。

依据 LCA 的分类过程可以看出, 测试样本邻域的定义是否准确直接影响分类器的选取合适与否。LCA 的基本准则是通过比较各个子分类器对于测试样本附近某一区域的分类准确率来衡量其正确识别该样本标签的能力。而测试样本邻域的正确界定是这种评估方式合理且准确的一个重要前提。如前所述, 通常通过计算测试样本与训练集或验证集中所有样本的距离来确定其邻域, 所以距离计算时所使用的距离度量方式会影响邻域的定义。因此, 在这个过程中如何选择合适的距离度量方式是一个需要进一步探索的问题。

2 邻域成分分析

NCA 算法的本质是将样本的特征向量映射到另一个空间, 在这个空间里 KNN 分类器拥有最佳的分类性能。假设 $X = \{x_1, x_2, \dots, x_n\}$ 是有标签的训练集样本, 其中 $x_i \in R^D$, 其对应的标签为 $\{l_1, l_2, \dots, l_n\}$ 。如果用一个矩阵 A 来表示这种映射, 实际上 NCA 学习的就是如下马氏距离 d 的度量准则:

$$d(x_i, x_j) = (Ax_i - Ax_j)^T (Ax_i - Ax_j) \quad (2)$$

在转换后的空间里, 利用一种基于随机近邻分配的微分损失函数衡量 KNN 分类器的性能。具体来说, 对于样本 x_i , 其选择样本 x_j 作为其近邻并继承其标签的概率是 p_{ij} 。这个概率 p_{ij} 通过转换后样本之间的欧氏距离来定义:

$$p_{ij} = \begin{cases} \frac{\exp(-\|Ax_i - Ax_j\|^2)}{\sum_{k \neq i} \exp(-\|Ax_i - Ax_k\|^2)} & j \neq i \\ 0 & j = i \end{cases} \quad (3)$$

式中, x_k 表示训练集 X 中除了 x_i 以外的任一样本。在随机选择的准则下, 可以计算 x_i 被正确分类的概率 p_i :

$$p_i = \sum_{j \in L_i} p_{ij} \quad (4)$$

式中, L_i 表示训练集 X 中标签与 x_i 的标签 l_i 相同的样本的下标的集合, 即 $L_i = \{j | l_j = l_i\}$ 。

在此基础上, 通过最大化正确分类样本的数量来学习得到转换矩阵 A , 其目标函数如下:

$$f(A) = \sum_{i=1}^n p_i = \sum_{i=1}^n \sum_{j \in L_i} p_{ij} \quad (5)$$

通过对 $f(A)$ 进行关于 A 的微分, 则可以获得如下梯度准则进行训练:

$$\frac{\partial f(A)}{\partial A} = -2A \sum_{i=1}^n \sum_{j \in L_i} p_{ij} (x_{ij} x_{ij}^T - \sum_{k=1}^n p_{ik} x_{ik} x_{ik}^T) \quad (6)$$

式中, $x_{ij} = x_i - x_j$ 。

3 方法的提出

由第2节可知, NCA 是一种度量学习方法, 其学习的目的是使 KNN 的分类准确率最高, 也就是说通过这样一个学习过程, 使每一个训练样本都能尽可能地找到它的真正近邻。采用 NCA 对训练数据进行处理后, 将会得到一个转换矩阵 A 。通过这样一个矩阵, 训练数据将会被转换到一个新的空间, 在这个空间里, KNN 能够获得较高的准确率。当对一个未知标签的测试样本 t 进行分类时, 首先需要通过矩阵 A 将其特征向量进行转换, 然后利用 KNN 在转换后的空间内寻找其近邻, 进而对其进行分类。通过这样一种方式, 测试样本被正确分类的概率增加, 也就是说在转换后的空间内, 更有可能找到它的真正近邻。而在 LCA 算法中, 测试样本 t 的邻域定义得越准确, 就越能确定一个更合适的局部域 ψ , 分类器在 ψ 上的分类准确率就更能反映其正确识别测试样本标签的能力, 因而能够选取出更适合 t 的子分类器。

基于以上思考, 将 NCA 算法引入 LCA 中, 用于定义测试样本 t 的邻域, 提出一种改进的 LCA 算法, 称之为 NLCA, 从而使得 LCA 能够更加准确地衡量每一个子分类器对测试样本 t 进行正确识别的潜力, 进而选取更加合适的子分类器对其进行处理。该方法实现流程如图 1 所示, 其实现步骤如下所述。

(1) 输入标签种类 Ω 、子分类器集合 C 、训练集 X 和测试集 T 、邻域尺寸 K 。

(2) 基于训练数据 X 利用 NCA 算法学习转换矩阵 A , 然后对训练数据进行转换, 得到转换后的训练数据 $X' = AX$ 。

(3) 采用如下步骤定义每个测试样本的邻域:

1) 利用转换矩阵 A 对测试样本 t 进行转换, 得到转换结果: $t' = At$; 2) 利用欧式距离计算 t' 与 X' 中所有样本的距离; 3) 找到 X' 中 t' 的 K 近邻, 用 N_K 表示这些近邻; 4) 定义 N_K 的原始样本为测试样本 t 的 K 近邻。

(4) 针对每一个测试样本 t , 采用所有子分类器对其标签进行预测, 如果所有子分类器输出的预测标签相同, 则将该标签作为该样本的类别。反之则通过以下步骤确定其标签。首先, 基于步骤(2)和步骤(3)确定的 K 近邻, 通过第 1 节所述 LCA 算法计算各个子分类器的局部精度 $LCA(m, j)$, ($m =$

$1, 2, \dots, M$), M 为分类器集合中包含的子分类器个数, 在分类器集合生成的时候确定。通常 M 可以根据经验大致设定一个值, 如果需要得到一个最优的 M 值, 可以研究子分类器个数对识别精度的影响, 识别精度最高时对应的子分类器数量即为最优 M 值。最后, 选择 $LCA(m, j)$ 值最大的子分类器对测试样本 t 进行标签预测。

通过上述步骤, 测试样本 t 的 K 近邻能够被识别得更准确, 有助于 LCA 算法定义一个更合理的局部域 ψ 。因而每个子分类器在 ψ 上的分类准确率与其正确识别测试样本 t 的潜力越相关, 后续通过 LCA 算法比较各个子分类器在 ψ 上的局部精度越能够选取出更合适的子分类器来识别测试样本 t 的标签, 最终提高分类性能。

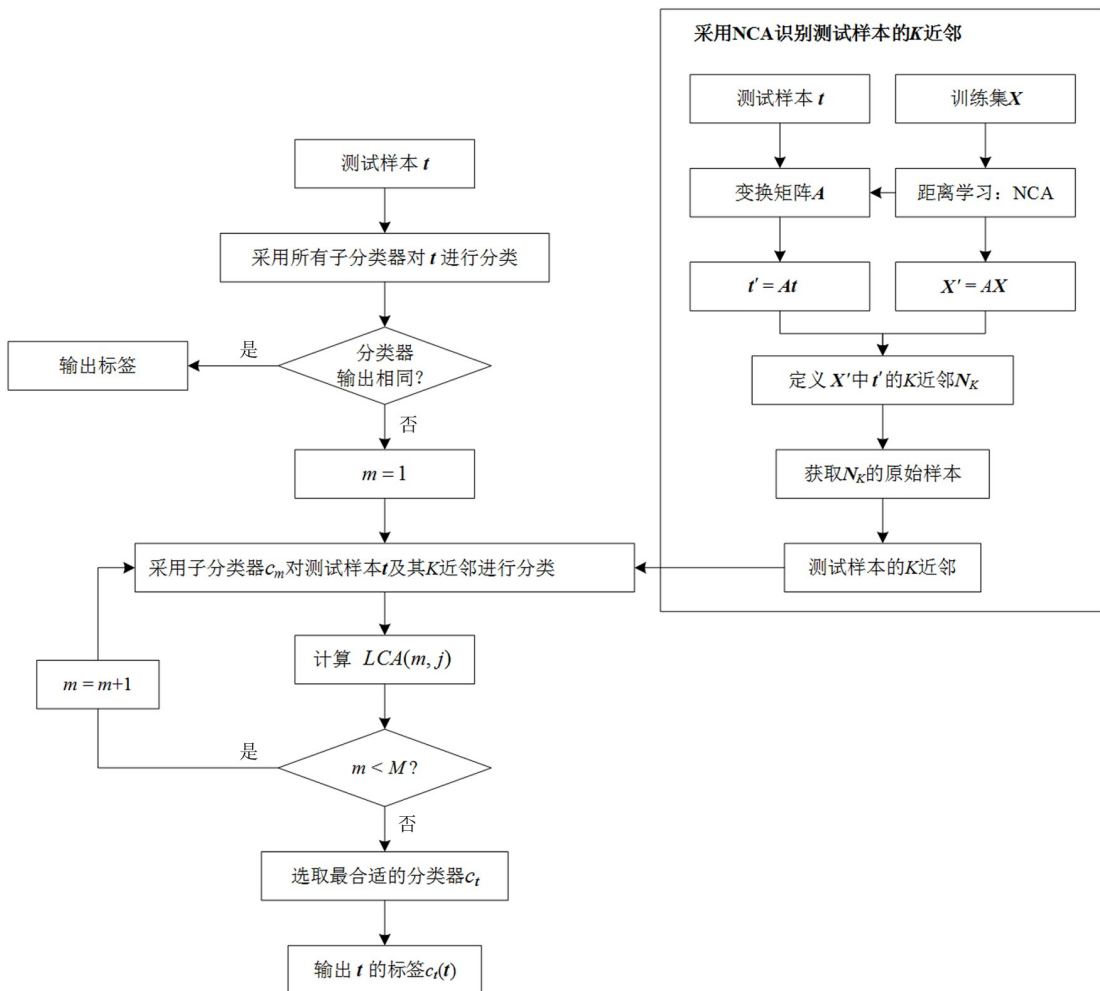


图1 NLCA 实现流程图

4 算法验证

为了验证 NLCA 在分类器动态选择中的有效性,并提高多分类器系统的分类准确率,采用 UCI 机器学习库^[19]中的标准数据对本文所提方法进行验证。因为本研究主要针对多分类问题,因此主要选取多分类问题数据进行研究。对于大部分数据,选择 50% 的数据作为训练集,其余 50% 的数据作为测试集,小部分数据根据实际情况进行调整。各组数据的基本情况如表 1 所示,第 1 列为数据集列表,第 2 列为各数据集的类别数,第 3~5 列分别为各数据集对应的训练样本数量、测试样本数量及特征维数。训练过程中,采用反向传播(back propagation, BP)神经网络^[20]作为子分类器算法,通过 AdaBoost 算法^[21]生成 10 个侧重于不同训练样本分类问题的 BP 神经网络,构成分类器集合。

表 1 UCI 数据集

数据集	类别数	训练集	测试集	特征维数
Wisconsin breast cancer	2	284	284	30
Vote	2	217	217	16
Wine	3	60	58	14
Balance	3	312	312	4
Iris	3	75	75	4
DNA	3	1 000	1 000	180
Vehicle	4	423	423	18
Ecoli	5	216	111	7
Image segmentation	7	1155	1155	18

表 2 列出了不同分类方法得到的分类结果,第 1 列为数据集列表,第 2 列为 10 个子分类器中准确率最高的 BP 神经网络对应的分类精度,第 3 列为融合 10 个子分类器预测结果得到的识别准确率,第 4 和 5 列分别为采用 LCA 和 NLCA 方法进行分类器动态选择、针对每个测试样本选择最合适的子分类器进行标签预测得到的识别精度。在 LCA 和 NLCA 方法中,选用的近邻数 $K = 1$ 。

由表 2 中的结果可知,对于大多数数据而言,采用 AdaBoost 算法集成多个 BP 神经网络能够显著提

表 2 分类准确率对比

数据集	单个 BP	AdaBoost	LCA	NLCA
Wisconsin breast cancer	86.97%	95.42%	88.73%	95.77%
Vote	92.17%	89.86%	91.78%	92.63%
Wine	75.86%	81.03%	84.48%	100.00%
Balance	86.54%	84.29%	84.94%	91.99%
Iris	85.33%	91.67%	93.33%	96.00%
DNA	75.10%	88.30%	77.60%	93.30%
Vehicle	68.79%	85.41%	71.69%	74.94%
Ecoli	75.68%	78.38%	80.18%	76.58%
Image segmentation	69.35%	72.99%	82.34%	87.19%

高分类精度,性能优于单个 BP 神经网络分类器。同时,采用 LCA 进行分类器动态选择能够提高其分类结果的准确性,但结果并不稳定,对于有些数据反而会降低其分类准确度。这是采用 K 近邻准则选择其邻域时造成的误差,因为缺乏一种有效的距离衡量方式,导致邻域的选择不准确。采用 NCA 算法对 LCA 进行改进后,NLCA 的分类精度得到了提高,同时也更加稳定。对于大部分数据都能取得比 AdaBoost 以及 LCA 更好的分类结果。因此,NLCA 算法能够解决 LCA 算法中的距离选择问题,通过选择更为准确的邻域,提高 LCA 算法的性能,从而能够根据测试样本从分类器集合中选取更为合适的子分类器对其标签进行预测,最终提高多分类器系统的分类准确度。

5 滚珠丝杠副状态识别实例

为了验证所提算法在滚珠丝杠副状态监测中的有效性与优越性,采用实验中获取的滚珠丝杠副状态监测数据进行状态识别验证。

5.1 实验数据获取

通常,滚珠丝杠副具有较高的可靠性,因此,正常情况下其寿命较长,需要很长时间才能采集到其失效数据。为了在较短时间获得足够的状态监测数据,采用文献[22]中如图 2 所示的滚珠丝杠副加速性能退化实验系统,开展滚珠丝杠副加速寿命实验。该实验系统通过增加滚珠丝杠副负载,加速其性能

退化,使其在较短时间失效。为了获得滚珠丝杠副性能退化过程中的状态监测数据,在丝杠螺母上安装了一个加速度传感器,采集其振动信号。加速寿命实验中,从滚珠丝杠副正常状态开始采集其状态监测信号,采样频率为 5 kHz,每隔 30 min 采集一次信号,直至滚珠丝杠副失效。



图2 滚珠丝杠副加速寿命实验系统

5.2 退化状态识别

在滚珠丝杠副的全寿命周期中,经历了正常、退化以及严重退化或失效 3 个阶段。本研究中,将滚珠丝杠副的状态按照以上情况分为 3 个状态,并采用多分类器系统进行状态识别,实现框架如图 3 所示。其主要包括离线训练和在线监测 2 个部分。离线训练过程中,对不同状态下获取的滚珠丝杠副振动信号进行处理,获取表征其退化状态的特征,利用 AdaBoost 算法训练 BP 神经网络集合。在线监测过程中,对于实时获取的信号特征,采用 NLCA 算法从分类器集合中选取最合适的子分类器进行状态识别,最终获得滚珠丝杠副的状态识别结果。

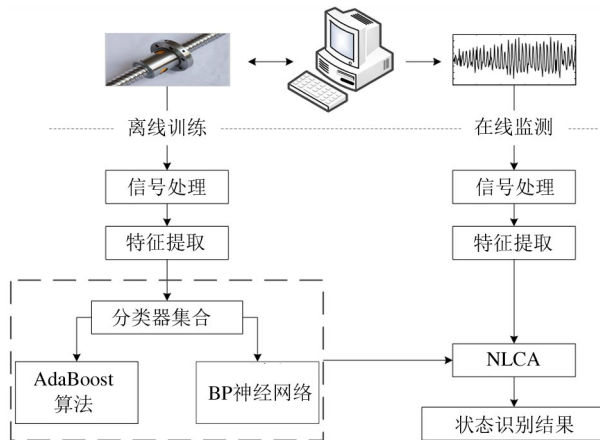


图3 基于 NLCA 的滚珠丝杠副状态识别框架

针对每组数据提取 10 组时域特征、28 组时频域特征以及 2 组三角函数特征作为其特征向量,如表 3 所示。其中,时频域特征包括采用小波包变换对信号进行 3 层分解得到的 8 个小波包系数的能量与能量比,即 8 个小波包频带的能量与能量比。同时还有利用经验模态分解对信号进行处理得到的前 6 个本征模态函数 (intrinsic mode function, IMF) 的瞬时能量与瞬时能量比,其计算公式分别如式 (7) 和 (8) 所示^[23]。由式 (7) 可知,IMF 瞬时能量不仅包含了 IMF 能量的大小,同时也能反映 IMF 沿时间尺度上的能量分布变化,因而比 IMF 能量更能表征信号的能量分布,更适合作为退化特征。

$$E_i = \sum_{k=1}^N (k \cdot \Delta t) (f_i(k \cdot \Delta t))^2 \quad (7)$$

式中, N 为信号样本点数, Δt 是采样点之间的间隔时间, $f_i (i = 1, 2, \dots, 6)$ 为第 i 个本征模态函数。将各个 IMF 的瞬时能量除以所有 IMF 瞬时能量之和,即得到归一化后的瞬时能量比向量,如式 (8) 所示,构成滚珠丝杠副信号的第 33 ~ 38 特征。

$$E = (E_1, E_2, \dots, E_6) / \sum_{i=1}^6 E_i \quad (8)$$

表3 信号特征

类型	特征	
时域特征	Y1:标准差	Y6:峰值因子
	Y2:峰-峰值	Y7:脉冲因子
	Y3:平均绝对值	Y8:偏度
	Y4:均方根	Y9:峭度
	Y5:波形因子	Y10:裕度因子
时频域特征	Y11-Y18:8 个小波包频带的能量	
	Y19-Y26:8 个小波包频带的能量比	
	Y27-Y32:6 个本征模态函数的瞬时能量	
	Y33-Y38:6 个本征模态函数的瞬时能量比	
三角函数特征	Y39:反双曲线正弦标准差	
	Y40:反正切标准差	

此外,分别利用反双曲线正弦函数与反正切函数对原始信号进行变换,并求取 2 个变换后序列的标准差作为特征,分别称为反双曲线正弦标准差 (standard deviation of inverse hyperbolic sine, SD of IHS) 以及反正切标准差 (standard deviation of inverse tangent, SD of IT),其计算公式如式 (9) 所示。

$$\begin{cases} \text{SD of IHS} = \sigma(\log(x_k + \sqrt{x_k^2 + 1})) \\ \text{SD of IT} = \sigma(\frac{1}{2}\log(\frac{1+x_k}{1-x_k})) \end{cases} \quad (9)$$

式中, $x_k(k = 1, 2, \cdots, N)$ 为第 k 个采样点处的信号幅值。

实现过程共选取了 100 组正常状态样本、77 组退化状态样本以及 40 组严重退化或失效状态样本。其中, 每种状态使用 50% 的数据作为训练数据, 其余 50% 的数据作为测试数据。

基于训练数据, 采用 AdaBoost 算法训练完成 BP 神经网络集合后, 针对测试样本, 分别采用 LCA 与 NLCA 进行分类器动态选择, 得到测试样本的输出结果, 表 4 为不同方法的分类结果。由表可知, 10 个子分类器中, 分类效果最好的子分类器整体分类准确率为 81.84%, 集成算法 AdaBoost 的整体识别精度为 88.88%, LCA 算法的精度为 93.52%, NLCA 算法的精度为 97.22%。其中, NLCA 算法的整体识别精度最高, 说明所提方法能够改善多分类器系统的性能, 提高滚珠丝杠副性能退化的识别精度。

表 4 状态识别结果

方法	NLCA	LCA	AdaBoost	单个 BP
准确率/%	97.22	93.52	88.88	81.84

图 4 为不同算法对于滚珠丝杠副不同性能状态的识别准确率。由图可知, NLCA 算法对于每种状态的识别率均比较高, 而其他算法对于正常状态的识别率较高, 对于失效状态的识别率也能达到 90% 左右, 但对于退化状态的识别效果不理想。单个 BP 神经网络的识别率仅有 52.63%, 采用 AdaBoost 算法能够改善这一状态, 识别率达到 73.68%, 但仍不够理想。采用分类器动态选择方法能够进一步提高其准确率, LCA 算法的准确率为 84.21%, NLCA 的准确率能够达到 97.37%。通过以上分析可知, 通过 NCA 度量学习方法, NLCA 能够为每个测试样本选择更为合适的邻域, 从而选取更为合适的子分类器对其进行分类, 最终达到提高多分类器系统状态识别精度的目的。

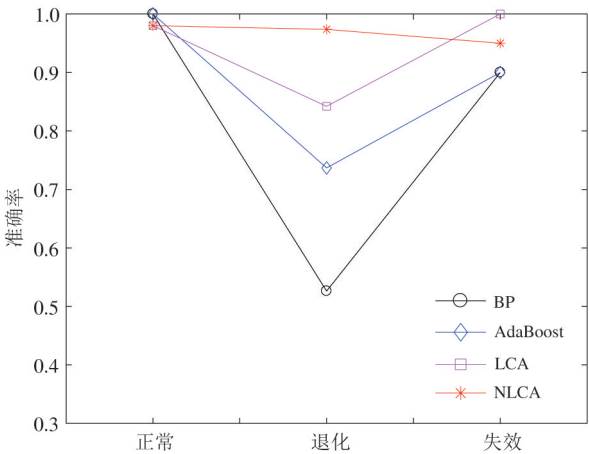


图 4 不同退化状态的识别精度

为了进一步说明所提方法的优点并探索滚珠丝杠副性能退化的本质特性, 将 3 种对比算法与 NLCA 算法的详细识别结果分别列于表 5 ~ 8 中。

对比这 4 个表格所列结果可知, 对于滚珠丝杠副性能状态识别而言, 正常状态最容易识别, 失效状态的识别也较为容易, 采用单个 BP 神经网络也能完成其识别任务。但是退化状态较难识别, 采用单

表 5 BP 神经网络识别结果

	正常	退化	失效	准确率
正常	100.00%	0.00%	0.00%	100.00%
退化	44.74%	52.63%	2.63%	52.63%
失效	5.00%	5.00%	90.00%	90.00%
准确率	100.00%	52.63%	90.00%	81.84%

表 6 AdaBoost 算法识别结果

	正常	退化	失效	准确率
正常	100.00%	0.00%	0.00%	100.00%
退化	26.32%	73.68%	0.00%	73.68%
失效	5.00%	5.00%	90.00%	90.00%
准确率	100.00%	73.68%	90.00%	88.88%

表 7 LCA 算法识别结果

	正常	退化	失效	准确率
正常	98.00%	2.00%	0.00%	98.00%
退化	13.16%	84.21%	2.63%	84.21%
失效	0.00%	0.00%	100.00%	100.00%
准确率	98.00%	84.21%	100.00%	93.52%

表 8 NLCA 算法识别结果

	正常	退化	失效	准确率
正常	98.00%	2.00%	0.00%	98.00%
退化	2.63%	97.37%	0.00%	97.37%
失效	0.00%	5.00%	95.00%	95.00%
准确率	98.00%	97.37%	95.00%	97.22%

一分类器、传统 AdaBoost 算法与传统分类器动态选择算法 LCA 均无法获得较为理想的结果。这可能是因为当滚珠丝杠副处于正常状态时,其工作较为稳定,因此振动信号也较为稳定,没有大的波动。同理,当滚珠丝杠副失效时,其振动信号也较为稳定。但是在退化过程中,随着退化程度的加深,滚珠丝杠副的状态存在较大波动,导致其振动信号存在一定程度上的变动,因此,对于退化状态的识别难度较大。

此外,由表 5~7 可知,退化状态容易被识别为正常状态,这是因为在滚珠丝杠副退化初期,由于退化程度较小,退化状态在振动信号上表现得并不明显,与正常状态的相似性大,因此,此时的退化状态易被错误识别为正常。采用 LCA 进行分类器动态选择能够在一定程度上提升退化状态的识别精度,但仍不够理想。这是因为使用的距离度量方式不合适,在对退化状态的数据样本进行标签预测时,将正常状态训练样本纳入待识别样本的邻域中,导致邻域定义不准确,进而影响 LCA 对各子分类器正确识别滚珠丝杠副未知状态潜力的准确评估,使得滚珠丝杠副的退化状态不能被准确地识别。而使用 NLCA 进行状态识别时,即便是退化状态样本,通过 NCA 算法也能找到其真正近邻,即标签为退化的训练样本,因而能够更加准确地评估各个子分类器对其进行正确分类的能力,从而提升退化状态的识别精度。

通过以上分析可知,采用所提方法能够比较准确地完成滚珠丝杠副的性能退化状态识别,NLCA 算法对于每种状态都具有较高的识别精度,性能优于传统 AdaBoost 算法和 LCA 算法。

5.3 子分类器数量的影响

如第 3 节所述,参数 M ,即分类器集合中子分类器的数量会对提出方法 NLCA 的分类结果有一定

的影响,通常 M 的值可以通过经验大致确定;但如果对精度要求比较高,可以通过设置不同的 M 值进行验证,从而得到一个最佳的子分类器数量值。图 5 为 $3 \leq M \leq 20$ 时,本文方法对于滚珠丝杠副状态的识别准确率。由图可知,当 $3 \leq M \leq 10$ 时,分类精度随着子分类器数量 M 的增加而增加,这是因为随着子分类器数量的增加,提出方法的选择范围变大,从而更有可能为每个测试样本选到最合适的子分类器。而当 $M > 10$ 以后,识别精度不再上升,一直稳定在 97.22%。这是因为采用 AdaBoost 生成子分类器时,当子分类器数量 M 达到一定的数量后,再生成的子分类器与之前生成的分类器相似度较大,子分类器的多样性并没有得到提升,因而本质上没有提供更多的选择,导致状态识别精度没有随着 M 的增加而增加。因此, $M = 10$ 是一个最佳的子分类器数量,前述内容中本文在对滚珠丝杠副进行状态识别时选择的子分类器数量也是 10。

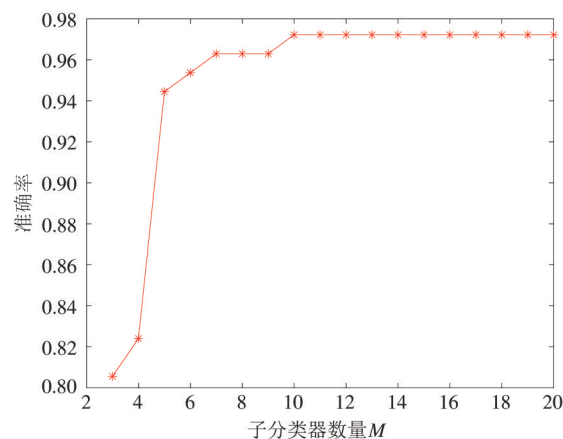


图 5 不同子分类器数量对滚珠丝杠副状态识别准确率的影响

6 结 论

(1) 针对多分类器系统中的分类器动态选择问题,提出了一种新的分类器动态选择算法 NLCA。研究表明,NLCA 的性能优于传统分类器动态选择方法 LCA,能够提高多分类器系统的分类准确率。

(2) 采用提出的 NLCA 方法成功实现了滚珠丝杠副的性能退化状态识别,达到了较为满意的精度。研究发现,滚珠丝杠副的正常与失效状态都较容易

被识别,而退化状态的识别存在一定难度。使用单个 BP 神经网络很难识别,采用传统 AdaBoost 算法、传统分类器动态选择方法也得不到较为理想的识别精度。在退化状态的识别过程中,前述方法容易将其识别为正常状态。采用 NLCA 方法能够很好地弥补这一缺陷,准确地识别退化状态,从而实现滚珠丝杠副性能状态的准确检测。

参考文献

- [1] TSAI P C, CHENG C C, HWANG Y C. Ball screw preload loss detection using ball pass frequency[J]. Mechanical Systems & Signal Processing, 2014,48(1-2): 77-91.
- [2] AZAMFAR M, LI X, LEE J. Intelligent ball screw fault diagnosis using a deep domain adaptation methodology[J]. Mechanism and Machine Theory, 2020, 151: 103932.
- [3] 李倩, 刘惠康, 皮瑶, 等. 基于深度置信网络的制动器故障诊断方法[J]. 高技术通讯, 2021,31(10): 1075-1080.
- [4] CHANG J L, CHAO J A, HUANG Y C, et al. Prognostic experiment for ball screw preload loss of machine tool through the Hilbert-Huang transform and multiscale entropy method[C]//IEEE International Conference on Information & Automation. Harbin, China:IEEE, 2010:376-380.
- [5] JIN W. Modeling of machine life using accelerated prognostics and health management and enhanced deep learning methodology[D]. Cincinnati: University of Cincinnati, 2016.
- [6] LI P, JIA X, FENG J, et al. Prognosability study of ball screw degradation using systematic methodology[J]. Mechanical Systems & Signal Processing, 2018,109(9):45-57.
- [7] 吴希曦, 高宏力, 燕继明, 等. 基于超球面支持向量机的丝杠故障诊断技术[J]. 计算机集成制造系统, 2010, 16(12): 2661-2667.
- [8] HUANG Y C, HSIEH Y K. Applying a support vector machine for hollow ball screw condition-based classification using feature extraction[J]. Proceedings of the Institution of Mechanical Engineers Part B Journal of Engineering Manufacture, 2020(3):095440542095884.
- [9] VAN M, HOANG D T, KANGH J. Bearing fault diagnosis using a particle swarm optimization-least squares wavelet support vector machine classifier[J]. Sensors, 2020, 20(12):3422.
- [10] 蔡艳平, 范宇, 陈万, 等. 改进时频分析和特征融合在内燃机故障诊断中的应用[J]. 中国机械工程, 2020,31(16):1901-1911.
- [11] CAO Y, SUN Y, XIEG, et al. A sound-based fault diagnosis method for railway point machines based on two-stage feature selection strategy and ensemble classifier[J]. IEEE Transactions on Intelligent Transportation Systems, 2022,23(8):12074-12083.
- [12] 卓鹏程, 夏唐斌, 郑美妹, 等. 基于 OHF Elman-AdaBoost 算法的滚动轴承故障多时期诊断方法[J]. 振动与冲击, 2021,40(6):71-78.
- [13] 吕雁飞, 侯子骄, 张凯. 多分类 BP-AdaBoost 算法研究与应用[J]. 高技术通讯, 2015,25(5):437-444.
- [14] WOODS K, KEGELMEYER W P, BOWYER K. Combination of multiple classifiers using local accuracy estimates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997,19(4):405-410.
- [15] KO A H R, SABOURIN R, BRITTO J A S. From dynamic classifier selection to dynamic ensemble selection[J]. Pattern Recognition, 2008,41(5):1718-1731.
- [16] IVAKHNENKO A. Heuristic self-organization in problems of engineering cybernetics[J]. Automatica, 1970,6(2): 207-219.
- [17] BRITTO A S, SABOURIN R, OLIVEIRA L E. Dynamic selection of classifiers: a comprehensive review[J]. Pattern Recognition, 2014,47(11):3665-3680.
- [18] GOLDBERGER J, HINTON G E, ROWEIS S, et al. Neighbourhood components analysis[J]. Advances in Neural Information Processing Systems, 2004,17: 513-520.
- [19] DUA D, GRAFFC. UCI machine learning repository [EB/OL]. [2023-07-24]. <http://archive.ics.uci.edu/ml>.
- [20] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors[J]. Nature, 1986,323(6088):533-536.
- [21] FREUND Y, SCHAPIRE R E. Experiments with a new boosting algorithm[C]//Proceedings of the 13th International Conference on International Conference on Machine

- Learning. Bari, Italy: Morgan Kaufmann Publishers Inc., 1996:148-156.
- [22] WEN J, GAO H L. Degradation assessment for the ball screw with variational autoencoder and kernel density estimation[J]. Advances in Mechanical Engineering, 2018, 10(9):168781401879726.
- [23] BIN G F, GAO J J, LIX J, et al. Early fault diagnosis of rotating machinery based on wavelet packets——empirical mode decomposition feature extraction and neural network [J]. Mechanical Systems & Signal Processing, 2012,27:696-711.

Ball screw condition recognition based on an improved dynamic classifier selection method

WEN Juan

(College of Mechanical Engineering, Zhejiang University of Technology, Hangzhou 310023)

(Hengfengtai Precision Machinery Co., Ltd., Wenzhou 325000)

Abstract

To improve the condition recognition rates of the ball screw, an improved dynamic classifier selection method is proposed. In this method, with neighborhood components analysis(NCA), the neighborhood of the test sample is defined accurately and adaptively without selecting the distance metric, and then the competence of each classifier in the multiple classifier system for accurately recognizing the testing task can be measured more exactly. Consequently, the classification accuracy is no longer restricted by the distance metric selection. The presented approach is applied to identify the health state of the ball screw. First, the AdaBoost algorithm is employed to create a back propagation (BP) neural networks pool. Then, to enhance the classification rates, the proposed dynamic classifier selection methodology is utilized to select the most suited classifier from the classifier pool for condition recognition according to the features extracted from the online signal. Experimental results show that the proposed method can identify the ball screw condition effectively with an accuracy of 97.22%, which is higher than that of the BP neural networks, AdaBoost, and the conventional dynamic classifier selection method.

Key words: dynamic classifier selection, neighborhood components analysis(NCA), condition recognition, ball screw, multiple classifier system