

基于 PCA 和 KNN 的电主轴故障诊断方法研究

黄国荣^a, 何亚飞^b, 杭琦^a

(上海第二工业大学 a. 环境与材料工程学院; b. 智能制造与控制工程学院, 上海 201209)

摘要: 电主轴是数控机床中重要的部件之一, 其性能的优劣直接影响机床工况和加工零件质量。对电主轴进行故障诊断能很大程度上提高数控机床的加工精度, 并且能够有效地增加其可靠性和安全性。在一般诊断过程中, 原始数据的高维特征量处理较为困难。为顺应实际应用中对电主轴故障诊断的精度要求, 提出一种基于主成分分析 (PCA) 与 K 最近邻 (KNN) 的电主轴故障诊断方法。此方法利用 PCA 对原始非线性时间序列数据的特征向量进行降维, 并选取其中主成分特征向量。将得到的主成分特征向量作为 KNN 的输入进行故障分类。最后将该方法的预测结果与决策树和随机森林的分类结果进行对比, 结果表明, PCA-KNN 算法在故障分类精度上相较于其他两种算法有显著提高, 是一种有效的电主轴故障分类方法。

关键词: 电主轴; 故障诊断; 主成分分析; K 最近邻

中图分类号: X823

文献标志码: A

Research on Fault Diagnosis Method of Motorized Spindle Based on PCA and KNN

HUANG Guorong^a, HE Yafei^b, HANG Qi^a

(a. Institute of Environmental and Material Engineering; b. Institute of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai 201209, China)

Abstract: Motorized spindle has become the most important part of modern numerical control. The performance of motorized spindle can directly affect the efficiency of computerized numerical control (CNC). Diagnosing when the potential failure occurs can highly extend CNC life, and increase its reliability and safety. But during the practical diagnosis process, it meets so many difficulties to handle high dimensional eigenvalues of tremendous amount of original collected data. To fulfill the demand of accurate classification of spindle failures, this paper reaches out a solution which based on principal component analysis (PCA) and K nearest neighbor (KNN) for bearings failure classification. This solution is first using PCA to reduce dimensionality of eigenvector of original data to select the principal component eigenvector and output the non-linear time series data. Set the outcome of PCA dimensionality reduction-the non-linear time series data-as the input of KNN, thus to get the bearings failure classification. And comparing to the results that come from decision tree and SVM, the conclusion shows that the accuracy of KNN combined with PCA is highly above decision tree and SVM, which indicates that the method in this paper is an efficient way of spindle failure diagnosis method.

Keywords: motorized; malfunction diagnosis; principal component analysis (PCA); K nearest neighbor (KNN)

收稿日期: 2018-07-24

通信作者: 何亚飞 (1960-), 男, 上海人, 教授, 硕士, 主要研究方向为智能制造、电主轴故障分析、数控机床机电一体化。

E-mail: yfhe@sspu.edu.cn

基金项目: 上海第二工业大学研究生教育、科研项目 (EGD17YJ0060) 资助

0 引言

随着当代制造业的发展和科技水平的提高,机械设备正朝着系统化和精密化的方向发展。与此同时,机械结构的复杂度急剧增大,自动化程度不断增高,这使得各个零部件之间的功能衔接也越为紧密^[1],导致故障诊断及故障修复的难度大大增加。伴随着机器学习、人工智能等信息技术快速发展,机械设备的故障诊断技术已经从传统的故障后检测向预测性检测方向发展。数控机床则是现代制造业的基础,电主轴是其核心的部件之一。相对于传统主轴,电主轴具有结构简洁、方便安装、转速高、惯量小、响应快、振动小、精度高等诸多优点,随着技术水平不断成熟和提高,电主轴已经迅速得到普及^[2]。不过在生产和实践中逐渐发现,电主轴在长期的高负荷的工作状态下,极易发生故障。其中关键部位如轴承、弹簧等微小损伤性故障若不能及时检测排出,就可能会引起连锁反应,轻则影响加工性能,重则造成整个系统的瘫痪。因此,对电主轴的故障趋势进行分析和预测对企业的生产工作将有着很大影响。

在以往的电主轴故障预测方法中,例如湖南大学的谭琛^[3]首先对各种典型的电主轴故障进行实验,采集相应的振动信号,然后采用小波变换等方式分析这些振动信号进而诊断故障。这种处理方式参数、方法较为单一,而且不能很好预测故障发生的时间。另外,采集到的振动信号基本都是非线性关系的,故对于此类问题,现通常采用机器学习算法来处理。例如浙江大学的赵月南^[4]提出的基于贝叶斯网络的电机故障诊断方法。但贝叶斯算法在训练的过程不能较好地处理连续值的属性,易产生过拟合或欠拟合的情况。另外,电子科技大学的李波^[5]和Intel产品有限公司的赵洁合作提出了将设备故障停机时间转换为设备不可用度、通过异常点替代和数据平稳化等两种数据预处理、建立零均值平稳随机序列进行ARMA建模。但是在模型调参的过程中问题会比较复杂。

在实际的故障趋势预测中,往往具有多组物理量^[6],针对每一组的非线性的时间序列数据,都可以提取很多有量纲和无量纲的特征向量进行故障的预测,而在将这些特征向量作为故障预测算法输入属性变量前,需要去除总冗余及干扰性较强的向

量,过多的向量会妨碍模型的建立,同时也会降低运算精度。故在某些应用场景预测之前,先会采用主成分分析(PCA)、核主成分分析(KPCA)等特征降维与特征融合方法^[7]对数据进行降维处理以显著提高预测精度。接着,会选取例如决策树、随机森林、 K 最近邻(KNN)等机器学习算法进行建模。其中,KNN算法是以非线性映射作为它的基础,具有很强的鲁棒性,对数据集的泛化能力强。既可以做分类模型也可以做回归模型。本文使用轴承退化过程中产生的非线性的时间序列数据,采用PCA对电主轴实验数据进行降维,选用KNN算法进行建模,最后,选取决策树、随机森林模型作为参照组进行故障诊断预测结果的效果比较。

1 PCA降维处理

在电主轴轴承故障趋势预测中,由于环境噪声和设备温度的影响,在传感器记录监测数据的过程中,数据一般都具有噪声。所以,对测量出来的时间序列物理量进行直接处理会影响预测精度。同时,对于不同的时间序列物理量之间可能互相关联,存在多重共线性问题。因此,在不知道不同特征向量对于预测的准确性是否有影响的情况下,直接进行故障趋势预测所得到的结果往往不尽如人意。所以在没有大量先验知识的情况下,对特征向量进行降维处理将会得到较理想的效果。

PCA^[8]是较常用的降维方法,它是利用坐标变换的思想,将高维的数据通过线性变换映射到低维空间中,也就是将一组可能存在相关性的变量转换为一组各维度线性无关的变量。但是在降维的过程中并不是将高维的数据进行直接的删减,而是对高维的特征向量进行了整合,计算出其协方差的特征值和特征向量。计算出的特征向量是经过重新整合的全新的正交特征向量。经过降维的特征向量排除了原始各向量之间的共线性对预测算法精度的影响,同时也改善了在数据样本较少但数据维度很高情况下的算法精度。使用PCA方法进行降维的基本步骤如下:

(1) 计算协方差矩阵。将原始数据中的每个样本用一个向量表示,构成一个 $m \times n$ 的矩阵,其中 m 为样本的个数, n 为原始数据特征向量的个数,并求出此矩阵的协方差矩阵 Σ 。

(2) 计算得出协方差矩阵 Σ 的特征值与特征向

量。求解 Σ 的所有特征值 λ_i 。对于每个不为 0 的特征值 λ_i , 计算其对应的特征向量 ξ_i 。

(3) 计算各个特征值贡献率

$$\alpha_i = \lambda_j / \sum_{i=1}^D \lambda_i, \quad j, i = 1, 2, \dots \quad (1)$$

式中, D 表示值不为 0 的特征值个数。

(4) 计算累计贡献率, 选择主成分。主成分的选择是根据累积贡献率

$$G(r) = \sum_{i=1}^r \lambda_i / \sum_{j=1}^s \lambda_j, \quad i, j = 1, 2, \dots \quad (2)$$

来确定的。式中, s 表示所有特征值个数, 选取累计贡献率达到 85% 以上的前 r 位特征值即为主成分。

2 KNN 算法

Cover 等^[9]于 1967 年提出 KNN 算法, 它是一种懒惰学习方法, 根据临近点类型判断样本点的类别。该方法的思路是: 首先计算待分类样本与已知训练样本的欧氏距离或余弦相似度, 找到与待分类文本距离最近或者相似度最大的 K 个最近邻文本。再根据 K 个最近邻文本的类别来判断待分类文本的类别^[10]。简单地说, 即 K 个最近邻文本中大多数属于某个类别, 则样本也属于这个类别^[11], 故在 K 值的选取上至关重要。如果 K 值偏小, 近邻数个数就会减少, 就会提高噪声的干扰从而降低分类精度; 如果 K 值偏大, 并且测试样本中属于训练集中包含数据较少的类, 则会增加噪声, 降低分类效果。选用偏小的 K 值, 就相当于用较小邻域里的训练集进行预测。极端状况下, 若 $K = 1$, 如果测试集只和最近的一个训练集有关, 误差会降低到最小。但如果此训练集正好是噪声, 那误差将会很大, 故很容易产生过拟合的现象。KNN 算法是一种简单高效的分类算法, 该算法规则分为以下 3 个步骤:

(1) 计算已经类别的数据集中的点与当前测试样本中点之间的距离, 并且按照距离递增的顺序进行排序;

(2) 选取与当前测试样本点距离最小的 K 个点;

(3) 统计前 K 个点所在类别中所出现的频率, 并返回前 K 个点出现频率最高的类别作为当前的预测分类结果。

3 基于 KNN 算法的故障诊断

3.1 实验数据

针对 KNN 算法在轴承诊断中的应用, 采用美国宇航局网站下载的数据^[12], 此数据是美国辛辛那提大学智能系统维护中心提供的滚动轴承全寿命周期的实验数据。此数据提取的加速度时间序列, 信号每隔 10 min 进行一次采样, 采样的频率为 20 kHz。本文采用的数据来源于其中一组实验, 开始于 2004-02-12 上午 10:32:39, 结束于 2004-02-19 早上 06:22:39, 数据记录了轴承从完好到退化的全生命周期过程, 样本总数为 984 组。

3.2 特征提取与 PCA 降维

由于监测的数据较多, 并且具有一定的噪声干扰, 故需要对原始数据提取特征量, 系统设备的监测和故障诊断常常使用时域信号的幅值参数。为了建立基于 KNN 算法的电主轴故障诊断模型输入输出之间的关系, 本文提取如下幅值参数特征, 其中 N 表示监测数据样本总数, x_i 表示样本点。

峭度指标

$$C_q = \frac{\frac{1}{N} \sum_{i=1}^N (|x_i| - \bar{x})^4}{X_{rms}^4} \quad (3)$$

峰度指标

$$I_p = \frac{X_p}{X_{rms}} \quad (4)$$

脉冲指标

$$C_f = \frac{X_p}{\bar{X}} \quad (5)$$

歪度指标

$$C_w = \frac{\frac{1}{N} \sum_{i=1}^N (|x_i| - \bar{x})^3}{X_{rms}^3} \quad (6)$$

峭度

$$\beta_q = \frac{1}{N} \sum_{i=1}^N (|x_i| - \bar{x})^4 \quad (7)$$

歪度

$$\beta_w = \frac{1}{N} \sum_{i=1}^N (|x_i| - \bar{x})^3 \quad (8)$$

平均值

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i(t) \quad (9)$$

峰值

$$X_p = \frac{\sum_{i=10} \max\{\text{样本}\}}{10} \tag{10}$$

有效值

$$X_{rms}^2 = \frac{1}{N} \sum_{i=1}^N x_i^2(t) \tag{11}$$

计算后得到以上参数指标部分数据如表 1 所示,并给样本分别赋予相应的样本序号,共 984 组。
根据计算得出的参数数据,以参数大小为纵坐标、以样本序号数(按时间先后顺序排布,范围为 1~984)为横坐标作出参数变化图(见图 1)。观察各参数变化图不难发现,部分参数变化趋势大致相同。但其中,有些变化趋势较为明显,有些较为隐晦,故

表 1 特征向量数据(8 组)
Tab. 1 The feature vector data (8 sets)

样本序号	1 号轴承峭度指标	1 号轴承有效值	1 号轴承峰值指标	1 号轴承脉冲指标	1 号轴承歪度指标
1	4.610 862	0.005 502	6.120 330	44.527 280	2.149 834
2	3.881 546	0.005 682	5.147 086	150.096 100	1.817 003
3	3.732 770	0.005 810	6.598 471	202.497 000	1.774 174
4	4.360 180	0.006 197	7.723 216	267.063 500	1.851 110
5	3.812 987	0.006 158	4.982 520	162.655 200	1.797 198
6	3.595 021	0.006 133	5.605 593	270.470 800	1.739 154
7	3.613 104	0.006 183	5.188 697	253.976 800	1.740 505
8	3.641 769	0.006 048	5.336 050	204.770 300	1.757 710

样本序号	1 号轴承峭度	1 号轴承歪度	1 号轴承平均值	1 号轴承峰值
1	0.000 139 607	0.000 877 502	-0.010 195 996	0.454
2	0.000 125 339	0.000 778 335	-0.002 585 010	0.388
3	0.000 126 046	0.000 785 905	-0.002 483 984	0.503
4	0.000 167 466	0.000 903 126	-0.002 276 611	0.608
5	0.000 144 602	0.000 868 517	-0.002 403 857	0.391
6	0.000 135 230	0.000 835 347	-0.001 623 096	0.439
7	0.000 138 130	0.000 846 215	-0.001 606 445	0.408
8	0.000 133 237	0.000 826 860	-0.002 026 660	0.415

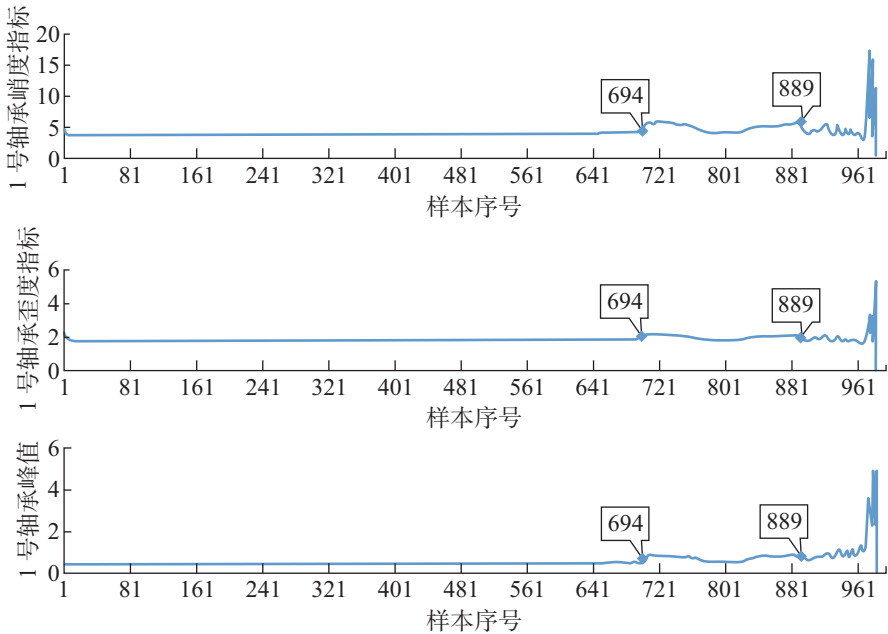


图 1 部分参数变化图
Fig. 1 Line charts of transformation of characteristics

存在数据冗余情况, 可以在变化趋势相同的参数中筛选出主要参数进行分析。为了降低数据冗余程度, 提高预测精度, 选取上述 9 个特征向量进行 PCA 并对高维的特征向量进行降维处理, 经过 PCA 降维得到分量的贡献率。按照 PCA 分量选取原则, 按照贡献率从上至下取值, 当取值的贡献率总和为 85% 左右时为合适, 作为 KNN 算法预测效果的检验。在本例中, 根据上述原则, 选取前 4 位分量, 总贡献率达 89.859%, 各主成分方差贡献率和起始特征值合计如表 2 所示。各特征向量和主成分因子载荷系数表(主成分矩阵)如表 3 所示。

表 2 部分特征分量贡献率
Tab. 2 Contribution rate of partial feature components

	特征变量贡献率			
	分量 1	分量 2	分量 3	分量 4
总贡献率/%	56.196	13.562	10.837	9.264
起始特征值合计	5.058	1.221	0.975	0.834

表 3 主成分矩阵
Tab. 3 Component matrix

1 号轴承	Component			
	1	2	3	4
峭度指标	0.888	-0.237	-0.031	0.313
有效值	0.948	0.098	-0.074	-0.119
峰值指标	0.399	-0.257	0.668	-0.394
脉冲指标	0.599	0.421	0.491	0.161
歪度指标	0.685	-0.384	-0.034	0.571
峭度	0.832	0.100	-0.398	-0.300
歪度	0.904	0.089	-0.302	-0.275
平均值	0.070	0.864	0.020	0.221
峰值	0.931	0.007	0.175	-0.014

在确定选取 4 个主成分之后, 需要确定主成分表达式以具体展现各项指标对于每个主成分的影响程度。在主成分表达式中, 变量前的系数的绝对值越大, 说明该指标对该主成分的影响程度越大。主成分表达式的系数矩阵中的每个元素的计算方法如式 (12) 所示; 主成分表达式系数矩阵如式 (13) 所示。式中: C_{ij} 表示第 i 个参数对第 j 个主成分矩阵; T_j 表示第 j 个主成分的起始特征值合计。由表 2、3 最后求得的主成分表达式分别如式 (14)~(17) 所示。

$$X_{ij} = \frac{C_{ij}}{\sqrt{T_j}}, \quad i, j = 1, 2, \dots \quad (12)$$

$$\begin{pmatrix} X_{11} & X_{21} & X_{31} & \cdots & X_{i1} \\ X_{12} & X_{22} & X_{32} & \cdots & X_{i2} \\ X_{13} & X_{23} & X_{33} & \cdots & X_{i3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{1j} & X_{2j} & X_{3j} & \cdots & X_{ij} \end{pmatrix} \quad (13)$$

$$F_1 = 0.394x_1 + 0.421x_2 + 0.177x_3 + 0.266x_4 + 0.304x_5 + 0.369x_6 + 0.401x_7 + 0.031x_8 + 0.413x_9 \quad (14)$$

$$F_2 = -0.214x_1 + 0.088x_2 - 0.232x_3 + 0.380x_4 - 0.347x_5 + 0.090x_6 + 0.080x_7 + 0.781x_8 + 0.006x_9 \quad (15)$$

$$F_3 = -0.031x_1 - 0.074x_2 + 0.676x_3 + 0.497x_4 - 0.034x_5 - 0.403x_6 - 0.305x_7 + 0.020x_8 + 0.177x_9 \quad (16)$$

$$F_4 = 0.342x_1 - 0.130x_2 - 0.431x_3 + 0.176x_4 + 0.625x_5 - 0.328x_6 - 0.301x_7 + 0.241x_8 - 0.015x_9 \quad (17)$$

3.3 实验方案及结果分析

3.3.1 KNN 预测模型的建立

首先, 确定训练集和预测数据。在分析参数变化图(见图 1)所展示的故障发生时参数的变化趋势可知, 在 1~694 范围内, 数据变化较小, 在 694 位置左右, 数据开始剧烈上升, 且在 694~889 有波动但是幅度不大, 在 889 处出现不规则剧烈振动, 并在少量延迟后快速上升。由于峭度指标、歪度指标、峰值指标数据变化趋势显著, 且符合实际情况中故障发生的过程, 故将这些参数所展现出的数值作为故障发生阶段的划分点。所以, 选取 694 这一点作为故障早期状态点, 选取 889 作为严重故障状态点。因此, 此样本集划分为 3 种状态: 1~694 的状态为正常, 695~889 为故障初始, 890~984 为损坏状态。

其次, 建立 KNN 分类模型关键需要选取合适的 K 值。 K 值的意义在于该数据是否被纳入分类中。过大的 K 值容易包含其他类别的值, K 值过小又容易受噪声点影响。对训练集建立 KNN 分类器, K 值

一般取奇数,这样不会在近邻中出现两类样本个数相同,从而不好确定待测点的类别现象。KNN 分类方法在不平衡数据中应用。为了选取最适用于该数据集的 K 值,选取 K 值为 3~60 范围内的奇数,再分别计算 30 个 K 值对应的准确率,并在 30 个 K 值对应的值分别画出散点图如图 2 所示。

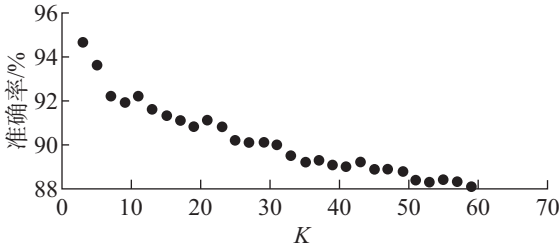


图 2 不同 K 值下的准确率
Fig. 2 Accuracy in different K values

准确率指标是评价分类器好坏的一个重要指标,数值越高分类效果越好,由图 2 可知,随着 K 值的增加,准确率逐渐下降, K 取 3 时准确率最大,因此 K 值取 3。本文中的 KNN 模型采用欧几里得距离计算两点的距离。

3.3.2 实验结果分析

为了验证本文采用的 KNN 的预测效果,选取随机森林算法和决策树算法对数据进行分类预测并比较其三者的分类效果,并对这 3 个模型做参数评估。

多类别的分类问题通常可以简化为二分类问题,机器学习领域中对于分类问题中常用的标准有: ROC 曲线分析^[13] 以及基于混淆矩阵^[14] (confusion matrix, 见表 4) 的准确率 (accuracy) 等。本文选择准

表 4 混淆矩阵
Tab. 4 Confusion matrix

Confusion		Predict		
		1	2	3
Real	1	a	b	c
	2	d	e	f
	3	g	h	i

准确率作为评价标准。准确率的计算公式为

$$\text{Accuracy} = \frac{a+e+i}{a+b+c+d+e+f+g+h+i} \quad (18)$$

根据混淆矩阵和准确率的计算方法分别得出 KNN 算法、决策树算法、随机森林算法分类模型的混淆矩阵如图 3 所示。根据之前所得结论,选择 KNN 算法 K 值为 3,并选择所有的数据作为测试集,可以得出 KNN 分类模型的准确率为 95.61%,而决策树算法的准确率为 84.55%,随机森林算法的准确率为 91.57%。各验证算法准确率结果如图 4 所示。由图可以看出,PCA-KNN 算法相比其他两种分类算法具有更好的分类精度,综合性能最强。

另外,通过 PCA-KNN 算法还可以对输入的参数进行分析,从而获得对模型判断故障阶段最具重要性的指标。本例中,PCA-KNN 模型参数重要性排序如图 5 所示。对比参数变化图,计算得出的重要性和实际相符合,变化越为显著、典型的参数对 PCA-KNN 模型重要性越大。



图 3 不同分类器下的混淆矩阵
Fig. 3 Confusion matrix under different classifiers



图 4 不同算法验证准确率
Fig. 4 Accuracy of different algorithms

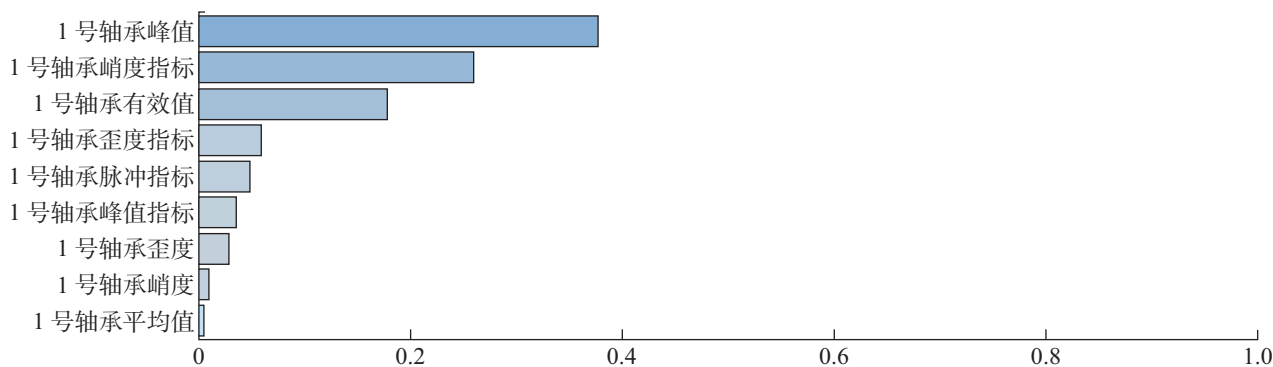


图5 不同参数对 PCA-KNN 模型判断的重要性排序

Fig. 5 The importance of different parameters for PCA-KNN model

4 结 论

本文提出了一种 PCA-KNN 算法用于提高电主轴故障诊断的分类精度。分析 PCA 降维与 KNN 算法的模型建立, 使用了美国辛辛那提大学智能系统维护中心提供的滚动轴承全寿命周期的实验数据进行了分类预测实验验证, 并使用了决策树算法和随机森林算法作为参照组进行实验结果对比分析, 利用混淆矩阵得到的准确率作为分类精度的评价指标, 证明了 PCA-KNN 算法的优越性 (该方法诊断准确率为 95.61%, 其余两种分别为 84.55% 和 91.57%)。但是在较大的数据量情况下, PCA-KNN 算法会对计算机造成比一般算法更大的计算压力, 预测时间也会明显增加, 在后期的研究中将对 PCA-KNN 算法进一步优化以提高运算速度和分类精度。

参考文献:

- [1] 时献江, 王桂荣, 司俊山. 机械故障诊断及典型案例解析 [M]. 化学工业出版社, 2013.
- [2] 田民. 高速机械密封试验机用电主轴设计与 CAD [D]. 合肥: 合肥工业大学, 2004.
- [3] 谭琛. 基于小波变换的电主轴振动故障检测技术研究 [D]. 长沙: 湖南大学, 2011.
- [4] 赵月南. 基于贝叶斯网络的电机故障诊断研究 [D]. 杭州: 浙江大学, 2016.
- [5] 李波, 赵洁, 郭晋. 设备故障评估新指标及基于 ARMA 的预测系统 [J]. 系统工程与电子技术, 2011, 33(1): 98-101.
- [6] RAUBER T W, BOLDT F D A, VAREJÃO F M. Heterogeneous feature models and feature selection applied to bearing fault diagnosis [J]. IEEE Transactions on Industrial Electronics, 2015, 62(1): 637-646.
- [7] 张恒, 赵荣珍. 故障特征选择与特征信息融合的加权 KPCA 方法研究 [J]. 振动与冲击, 2014, 33(9): 89-93.
- [8] TIPPING M E, BISHOP C M. Probabilistic principal component analysis [J]. Journal of the Royal Statistical Society, 1999, 61(3): 611-622.
- [9] COVER T M, HART P E. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.
- [10] 王煜. 基于决策树和 K 最近邻算法的文本分类研究 [D]. 天津: 天津大学, 2006.
- [11] 石欣, 印爱民, 张琦. 基于 K 最近邻分类的无线传感器网络定位算法 [J]. 仪器仪表学报, 2014(10): 2238-2247.
- [12] Acoustics and Vibration Database. IEEE PHM 2012 Data Challenge bearing dataset [Z/OL]. [2018-03-07]. <http://data-acoustics.com/measurements/bearing-faults/bearing-6/>.
- [13] FAWCETT T. An introduction to ROC analysis [J]. Pattern recognition letters, 2006, 27(8): 861-874.
- [14] 张岁霞, 严传波, 木拉提·哈米提, 等. KNN 分类器在新疆哈萨克族食管癌分型中的应用 [J]. 科技通报, 2016, 32(8): 46-50.