基于随机森林的航空发动机工作状态识别

李鼎哲，彭靖波，赵泽平，王玮轩，赵　彪
（空军工程大学航空工程学院，西安，710038）

摘要　为解决人工识别航空发动机工作状态中存在的误判和耗时费力等问题，提高识别准确率，提出了一种基于主成分分析(PCA)的特征提取方法和随机森林(RF)的智能识别方法。首先对飞参数据进行预处理，利用PCA将数据降维进行属性约简，并根据发动机工作状态将样本分组，用随机森林方法训练获得分类器；然后将几种分类方法的识别效果进行对比；最后采用该方法对某一架次的发动机工作状态进行识别。结果表明，该方法能够准确快速地识别航空发动机的稳定工作状态，识别准确率达到97.89%。可应用于发动机工作状态的相关研究。

关键词　航空发动机；飞参数据；工作状态识别；随机森林；主成分分析；属性约简
DOI　10.3969/j.issn.1009-3516.2020.01.003
中图分类号　V235.13　　**文献标志码**　A　　**文章编号**　1009-3516(2020)01-0015-06

An Aero-Engine Working Condition Recognition Based on Random Forest

LI Dingzhe, PENG Jingbo, ZHAO Zeping, WANG Weixuan, ZHAO Biao
(Aeronautics Engineering College, Air Force Engineering University, Xi'an 710038, China)

Abstract　In order to solve the misjudgment and time-consuming problem in manual identification of aero-engine working condition, and to improve the accuracy of the recognition, an intelligent recognition method based on principal component analysis(PCA)and a random forest(RF)is proposed. Firstly, PCA is used to reduce the dimensionality of the original flight data preprocessed, and the processed data on the basis of aero-engine working condition are grouped, and then RF are constructed. Secondly, the recognition effect of several classification methods is made in comparison with each other. At last, the method is used to recognize the working condition of one sort. The experiment results indicate that the recognition accuracy is 97.89% by this method. And this method is able to recognize the aero-engine working condition fast and accurately, and simultaneously is able to apply to the research related to the aero-engine working condition.

Key words　aero-engine; flight data; working condition recognition; random forest; principal component analysis; attribute reduction

航空发动机工作状态识别属于模式识别中的多分类问题。目前，已有学者将SVM与SVDD方法用于航空发动机工作状态识别，文献[1]基于最小二乘支持向量机(LS-SVM)将一对一、一对多以及纠

收稿日期：2019-09-07
基金项目：国家自然科学基金(51506221)；陕西省自然科学基础研究计划(2015JQ5179)
作者简介：李鼎哲(1996—)，男，浙江宁波人，硕士生，主要从事航空发动机故障诊断研究。E-mail:15306605060@163.com
通信作者：彭靖波(1980—)，男，湖南南县人，副教授，主要从事航空发动机分布式控制研究。E-mail:pjb1209@126.com

引用格式：李鼎哲，彭靖波，赵泽平，等.基于随机森林的航空发动机工作状态识别[J].空军工程大学学报(自然科学版)，2020，21(1)：15-20.LI Dingzhe，PENG Jingbo，Zhao Zeping，et al. An Aero-Engine Working Condition Recognition Based on Random Forest[J]. Journal of Air Force Engineering University (Natural Science Edition)，2020，21(1)：15-20.

错输出编码 3 种分类方法进行了比较,并采用纠错输出编码方法对某架次发动机工作状态进行了识别。但所提方法在追求分类速度的同时牺牲了一定的分类精度,并且数据缺失对分类性能有较大的影响。文献[2]构建了一种基于超椭球分类面支持向量数据描述(HE-SVDD)分类器,具备了快速从大规模飞行数据中识别航空发动机工作状态的能力。但所提方法的分类性能依赖于核函数的选取,且核函数的选取只能依靠经验。文献[3]针对 HE-SVDD 方法存在的部分缺陷进行改进,提出了一种改进 BA 优化的多核支持向量数据描述(CRBA-MKSVDD)分类算法,进一步提高分类器的性能。但所提方法作为一种单分类器,存在响应时间长等缺点。

随机森林(Random Forest, RF)作为一种统计学习理论,利用 Bootstrap 重抽样方法从原始样本中抽取多个样本,对每个样本建立决策树模型,然后组合多棵决策树的预测,通过投票得出最终预测结果。该方法内部执行交叉验证,对于复杂和非线性数据,有很好的预测效果,并且有训练速度快、不易过拟合等优点[4-5],近年来广泛应用于故障诊断[6-7]、聚类识别[8-9]、回归预测[10-11]等领域。PCA 法作为一种数据处理分析方法,主要应用于图形、语音等方面的处理和识别以及特征选择[12-14]。为此,本文将主成分分析法(Principal Component Analysis, PCA)与随机森林(RF)结合对航空发动机工作状态进行识别。

1 状态识别方法

1.1 主成分分析方法

PCA 是一种常用的数据分析方法,其原理是通过一个向量矩阵将原始数据从高维空间投影到一个低维的向量空间[15-16]。换言之即通过线性变换将原始数据变换为一组各维度线性无关的表示,以此提取数据的主要线性分量。PCA 法的流程为:①样本向量集;②计算矩阵 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{C}$;③计算协方差矩阵 $\boldsymbol{C}$ 的特征值和对应特征向量;④将所得特征向量从大到小排列对应的特征向量组成特征矩阵 $\boldsymbol{U}$;⑤使用特征矩阵 $\boldsymbol{U}$ 将样本特征矩阵 $\boldsymbol{X}$ 进行变换;⑥输出主成分。

设一个 n 维样本向量集 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$,则 $\boldsymbol{X}\subset\boldsymbol{R}^{m\times n}$,令:

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1}$$

得到样本集的协方差矩阵为:

$$\boldsymbol{C}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^{\mathrm{T}} \tag{2}$$

将矩阵 $\boldsymbol{C}$ 正交分解,得到:

$$\boldsymbol{C}=\boldsymbol{U}\cdot\boldsymbol{\Lambda}\cdot\boldsymbol{U}^{\mathrm{T}} \tag{3}$$

式中:$\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_n)$ 是对角阵,由 $\boldsymbol{C}$ 的 n 个按降序排列的特征值 λ_i 组成。特征矩阵 $\boldsymbol{U}=[u_1,u_2,\cdots,u_n]$ 由特征值 λ_i 对应的特征向量 $\boldsymbol{u}_i(i=1,2,\cdots,n)$ 组成的特征矩阵。λ_k 对应的贡献度为:

$$P_k=\lambda_k\left(\sum_{i=1}^{n}\lambda_i\right)^{-1} \tag{4}$$

为了提取样本集中信息量大的主元,用贡献率 θ 来表示,得到前 d 个主元的贡献率为:

$$\theta=\sum_{k=1}^{d}P_k=\sum_{k=1}^{d}\lambda_k/\sum_{i=1}^{n}\lambda_i \tag{5}$$

设定阈值为 P,使得 $\theta\geqslant P$,确定主元,可得到主元模型:

$$\boldsymbol{V}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{X} \tag{6}$$

原先的矩阵 $\boldsymbol{X}$ 可以重构为:

$$\dot{\boldsymbol{X}}=\sum_{i=1}^{d}\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}_i \tag{7}$$

这样就可以将前 d 个特征向量构成的 PCA 子空间的大部分特征信息体现出来,实现了属性约简的目的。

1.2 决策树

决策树(Decision Tree)[17]方法可认为是一棵分类模型树,包含根节点、内部节点和叶节点,图 1 为决策树的基本构成。

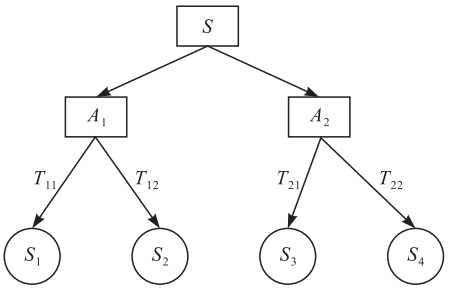


图 1 决策树基本构成图

其中,根节点包含整个数据集,每个内部节点是一个判断条件,它将根据判断条件的测试结果,将数据集分配到 2 个或多个子节点,子节点继续分裂直至产生叶节点,包含最终的数据类别。但决策树生长过渡会使其产生过拟合的问题,且对于不平衡样本的分类性能较差,信息增益容易偏向样本量大的特征。

1.3 随机森林算法

随机森林是由多棵决策树组成的组合分类器,

图 2 为随机森林的算法流程图。通过训练多个树状分类器，将多棵决策树的预测组合，最后经过投票的方式得到预测结果。其基本思想是先采用 Bootstrap 抽样从原始训练集中抽取 k 个样本，其次建立 k 个决策树模型，获得 k 种分类结果，最后对所有结果投票表决，确定最终归属于哪一类别。其模型函数为：

$$H(x) = \arg\max_{Y} \sum_{i=1}^{k} I(h_i(x) = Y) \tag{8}$$

式中：k 为决策树的数量；Y 为输出变量（目标变量）；I 为示性函数；$H(x)$ 表示组合分类模型；$h_i(x)$ 表示第 i 棵决策树的分类模型。

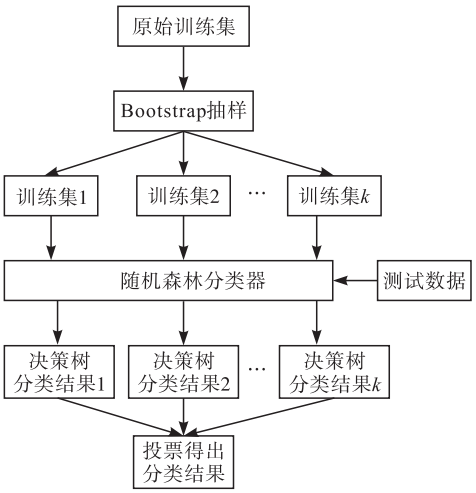


图 2　随机森林流程图

随机森林通过构造不同的训练集增加分类模型间的差异，从而提高组合分类模型的外推预测能力[1]。其随机性主要体现在以下方面：第一，训练样本选择具有随机性，即通过多次有放回抽样形成子集；第二，特征子集的选择具有随机性，即随机抽取特征集合；第三，所有决策树模型不进行剪枝，自由成长。因此，随机森林很好地解决了过拟合的问题，将多个弱分类器集成一个强分类器。

1.4　算法步骤

算法设计流程主要包含了某型发动机飞参数据的采集与预处理、特征提取以及工作状态识别。

首先，将相关发动机参数从飞参记录器转录至地面处理设备（通常是便携式计算机），进行数据的预处理，随后按一定比例选取训练集和测试集。再采用 PCA 方法对数据集进行特征提取，利用降维后的训练集对随机森林分类器进行训练，再导入测试集进行发动机工作状态的分类识别，并计算分类准确率和测试时间。

1）采集飞参数据，提取相关特征参数并进行预处理。

2）通过 PCA 方法将所提取的飞参特征数据进行降维，根据贡献率选择 n 个主成分，输出对应的特征向量矩阵，组成训练数据集。

3）在训练数据集中通过 Bootstrap 方法有放回抽取 k 个样本集，构建 k 棵决策树。

4）在每一棵树的各节点处随机抽取 m 个特征属性（$m \leqslant n$），对评估效果最佳的属性在对应节点处遵循节点不纯度原则进行分裂生长。

5）每棵决策树充分生长，不进行任何剪枝。

6）将生长得到的 k 棵树组成随机森林，根据分类器的投票数量得到相应分类结果。

上述算法设计流程如图 3 所示。

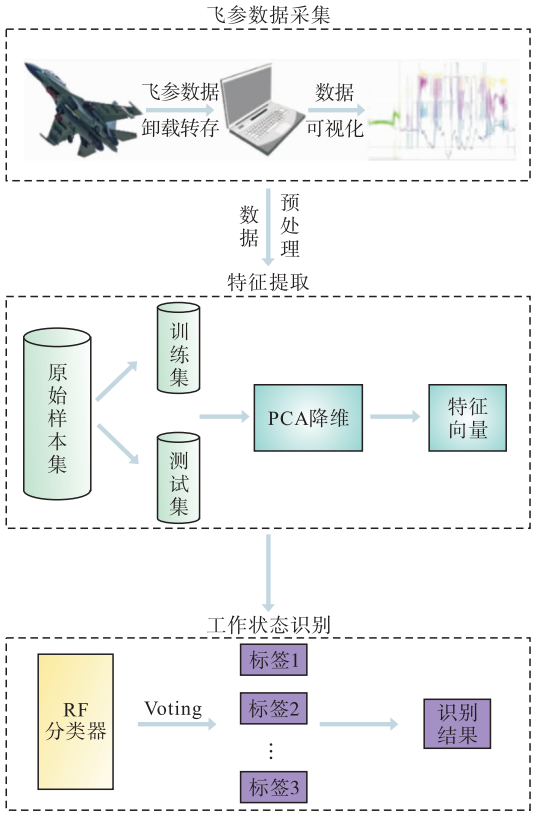


图 3　算法流程图

2　航空发动机工作状态识别

2.1　飞参数据选取与预处理

某型发动机的稳定工作状态包含慢车、节流、中间、小加力和全加力（最大）状态，在外场工作中，通常需要将油门杆角度与其他同发动机相关的参数结合起来人工判读发动机工作状态，因此在特征飞参数据的选取上将会以此作为参考。

以下原则将会在参数选取过程中被考虑：①以该型号发动机技术说明中明确规定的相关技术指标以及对应参数为准。②若飞参数据之间存在较强的

相关性,则选择相对工作状况强相关的参数,如换算转速与转速之间存在关联,考虑到转速作为发动机工作状态划分的主要依据之一(如慢车状态转速通常为中间状态转速的 0.4～0.6 倍),而换算转速更多的用于发动机相关参数的控制规律,那么就选择转速作为特征参数。

综上,最终选择油门杆角度(A_{PL},(°))、低压转速(n_1,%)、高压转速(n_2,%)、滑油压力(P_m,MPa)、主燃油量(W_f,kg)、涡轮后温度(T_6,℃)、涡轮后压力(P_6,kPa)、发动机排气温度(T_9,℃)、喷口面积(A_9,cm^2)以及加力接通信号(K)共计 10 个特征参数。

从外场收集该型航空发动机 2018 年 5 月日常飞行训练中的飞参数据。随机选中 4 个无故障飞行架次,对上述的特征参数进行提取,根据文献[18]所提方法进行如下预处理:

1)异常值剔除。对于明显偏离参数正常变化范围且同一时间点其余参数均正常的点,为避免影响分类效果,应当剔除。

2)同步性处理。某型飞机飞参记录器 1 s 记录 4 帧飞参数据,但由于不同的参数采样频率不同,在时间上并不同步,需要进行同步性处理,处理的办法是对各参数在 1 s 内求均值。

3)数据归一化。由于所选参数的测量精度以及量纲的不同,需要进行归一化处理,将所有参数归一化至 0～1 之间。

按照上述原则和处理方法最后得到原始样本数据 38 826 个,其中慢车、节流、中间、小加力、全加力数据数量分别为 10 416、9 892、12 208、2 398 和 3 912个。

2.2 属性约简和决策树数目选择

为降低特征维数以及减少各特征间相关性,采用 PCA 方法对选取的 10 个特征进行融合和约简。

5 个状态下的样本各取 70%作为训练集,余下 30%作为测试集。对所取训练集进行 PCA 处理,可以得到 10 个特征值矩阵 $\boldsymbol{\Lambda}$ 以及对应的特征向量 $\boldsymbol{U}$。选取主元累计贡献率 θ 为 95%,得到相应的 k 值为 5。前 6 个主元的累计贡献率分别为 59.1%,69.6%,79.4%,87.8%,95.2%,96.6%。

在进行状态识别前,需要选择最优的决策树数目。决策树数目与分类准确率的关系如图 4 所示。可以看到当决策树棵树为 15 时,分类准确率达到 98.43%,且随着决策树数目增多,准确率趋于稳定。但决策树增多会使计算复杂度随之上升,伴随着计算时间的增加。因此,选择 15 棵决策树组成随机森林分类器,进行发动机工作状态的识别,既能保证分类精度,又能合理的减小计算复杂度,缩短计算时间。

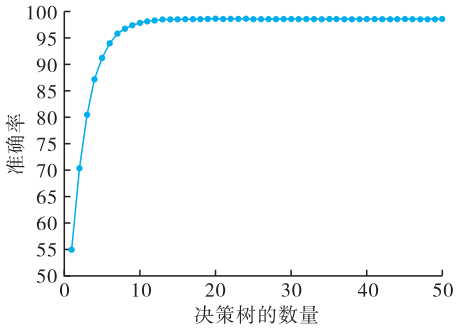


图 4 分类准确率与决策树数目关系图

表 1 比较了未使用和使用 PCA 方法进行属性约简后的随机森林分类器(决策树数量同为 15)分类准确率和训练时间。可以看出,对数据进行属性约简后,训练时间将会显著减少,而且分类精度仍然较高。

表 1 2 种方法准确率比较

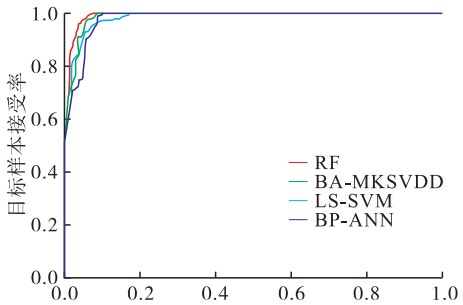
算法	准确率/%	训练时间/s
RF	97.48	29.33
PCA-RF	97.33	5.58

2.3 基于随机森林的训练与测试

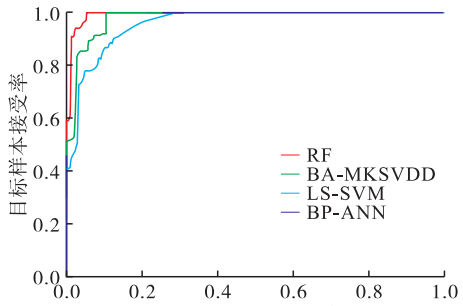
实验过程中,选择属性约简后的训练集对不同的分类器(BP-ANN、LS-SVM、BA-MKSVDD 和 RF)进行训练,用同样经过属性约简的测试集对训练后的分类器进行分类精度检验。图 5 为反映分类器识别效果的受试者工作特性(ROC)曲线。

对比分析图 5 可知,所提出的 PCA-RF 方法在发动机的 5 种工作状态下都具有比较优异的分类性能,相比于其它 3 种识别方法尤其是 BP 神经网络和 LS-SVM 而言,其对 5 种工作状态下的特征数据,在较低的异常样本接受率下都能够正确的接受大部分目标样本,更适合用作状态识别分类器。

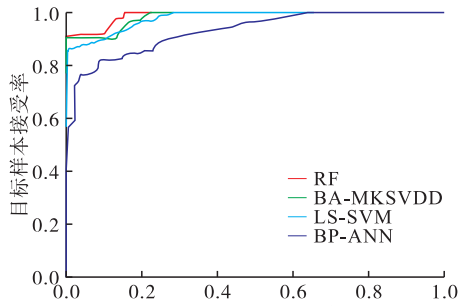
表 2 和表 3 分别为使用 PCA 降维前后 4 种分类器分类精度和测试时间。从表 2 可知,RF 的识别准确率最高,明显高于 LS-SVM 与 BP-ANN,尤其表现在发动机进入加力工作状态之前的 3 个工作状态上。由于发动机进入加力状态工作时间较少,以及加力状态下飞参数据具有波动性强、稳定性低的特点,因此造成识别准确率的下降。由表 3 可知,使用 PCA 降维后,能够显著减少识别时间,但同时会使识别准确率有小幅下降。综合看来,本文所选的 PCA-RF 方法既可以有效提高识别效率,又能够保证较高的识别精度。



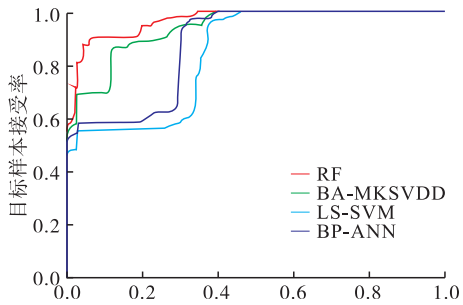
(a) 慢车状态



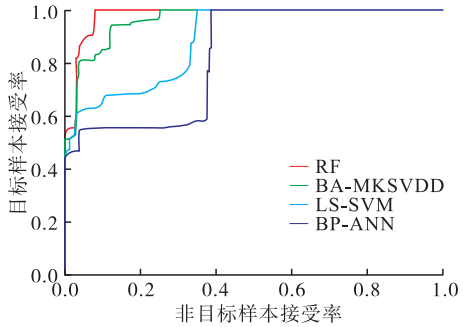
(b) 节流状态



(c) 中间状态



(d) 小加力状态



(e) 全加力状态

图 5　不同工作状态的 ROC 曲线

表 2　使用 PCA 降维前各分类器的分类精度与测试时间

分类器	测试时间/s	准确率/%				
		慢车	节流	中间	小加力	全加力
RF	9.30	99.27	98.98	99.46	93.87	95.82
BA-MKSVDD	10.46	98.06	98.18	98.14	93.36	95.10
LS-SVM	1.57	94.35	95.62	96.06	90.95	93.56
BP-ANN	2.76	93.46	95.39	96.03	91.02	93.38

表 3　使用 PCA 降维后各分类器的分类精度与测试时间

分类器	测试时间/s	准确率/%				
		慢车	节流	中间	小加力	全加力
PCA-RF	2.09	99.13	98.88	99.29	93.71	95.64
PCA-BA-MKSVDD	4.42	97.49	98.02	97.99	93.27	94.84
PCA-LS-SVM	1.06	94.27	95.56	95.91	90.73	93.44
PCA-BP-ANN	1.93	93.38	95.28	95.95	90.81	93.09

2.4　状态识别实例

使用本文提出的算法，节选该型发动机的某次飞行训练中的一段飞参数据进行工作状态识别，在进行发动机工作状态状态识别前需要利用 2.1 节中提出的原则对飞参数据进行预处理。

在选取的这段飞参数据内，该型发动机先后经历了慢车、节流、慢车、中间、节流、小加力、全加力、最大、节流和慢车状态，图 6 为识别结果。

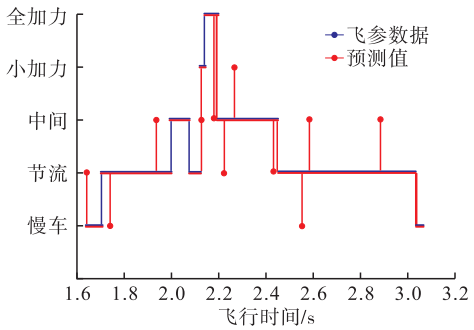


图 6　某架次发动机工作状态识别结果

可以看出，预测结果同实际结果吻合度较高。使用本文所提方法对该段发动机工作状态识别准确率达到 97.89%，已经基本符合发动机的实际工作状况，可以体现本文方法的有效性。

3　结论

本文提出了一种基于 PCA 的特征提取方法和 RF 的航空发动机工作状态识别方法。通过对某型发动机工作状态的识别实例，得出以下结论：

1）利用 PCA 方法进行属性约简对识别准确率影响较小，同时能提高识别效率。

2）经过对比实验，本文所提方法具有较高的识别准确率和识别效率。

3)节选某架次航空发动机飞参数据进行工作状态识别,结果表明本文所提方法对发动机工作状态能有效识别,具有研究应用价值。

此外,随机森林分类器的分类性能易受样本数量影响,对于小样本数据的分类效果仍有提高的空间。

参考文献

[1] 曲建岭,李晓娟,司敬国,等.基于飞参数据的发动机工作状态识别方法研究[C]//第32届中国控制会议论文集.西安:中国控制学会,2013:3565-3569.

[2] 周胜明,曲建岭,高峰,等.基于HE-SVDD的航空发动机工作状态识别[J].仪器仪表学报,2016(2):308-315.

[3] 何大伟,彭靖波,胡金海,等.改进BA优化的MKS-VDD航空发动机工作状态识别[J].北京航空航天大学学报,2018(10):2238-2246.

[4] LI B,WEI Y,DUAN H,et al. Discrimination of the Geographical Origin of Codonopsispilosula Using Near Infrared Diffuse Reflection Spectroscopy Coupled with Random Forests and *k*-Nearest Neighbor Methods[J]. Vibrational Spectroscopy,2012,62:17-22.

[5] RODRIGUEZ-GALIANO V F,GHIMIRE B,ROGAN J,et al. An Assessment of the Effectiveness of a Random Forest Classifier for Land-Cover Classification [J]. Journal of Photogrammetry and Remote Sensing,2012,67:93-104.

[6] 艾延廷,董欢,田晶,等.一种航空发动机中介轴承故障诊断方法[J].机械设计与制造,2018(10):157-160.

[7] CERRADA M,ZURITA G,CABRERA D,et al. Fault Diagnosis in Spur Gears Based on Genetic Algorithm and Random Forest[J]. Mechanical Systems and Signal Processing,2016,70:87-103.

[8] 方勇,龙啸,黄诚,等.基于LSTM与随机森林混合构架的钓鱼网站识别研究[J].工程科学与技术,2018(5):196-201.

[9] 周绮凤,杨小青,周青青,等.基于随机森林的建筑结构损伤识别方法[J].振动·测试与诊断,2012(2):197-201.

[10] PALMER D S,O'BOYLE N M,GLEN R C,et al. Random Forest Models to Predict Aqueous Solubility [J]. Journal of Chemical Information and modeling,2007,47(1):150-158.

[11] 方匡南,朱建平,谢邦昌.基于随机森林方法的基金收益率方向预测与交易策略研究[J].经济经纬,2010(2):61-65.

[12] 范群贞,刘金清.基于PCA/ICA的人脸特征提取新方法[J].电子测量技术,2010,33(8):31-34.

[13] TAKIGUCHI T,ARIKI Y. PCA-Based Speech Enhancement for Distorted Speech Recognition[J]. Journal of multimedia,2007,2(5):13-18.

[14] 黄梦莹,张晓滨.融合CHI与信息增益的情感文本特征选择[J].西安工程大学学报,2018,32(6):713-717.

[15] ZHANG D,ZHOU Z H,CHEN S. Diagonal Principal Component Analysis for Face Recognition[J]. Pattern recognition,2006,39(1):140-142.

[16] 胡帅,顾艳,姜华,等.基于PCA-BPNN的学生写作成绩预测模型研究[J].国外电子测量技术,2015,34(12):35-38.

[17] TSO G K F,YAU K K W. Predicting Electricity Energy Consumption:A Comparison of Regression Analysis, Decision Tree and Neural networks[J]. Energy,2007,32(9):1761-1768.

[18] ASHKEZARI A D,MA H,SAHA T K,et al. Application of Fuzzy Support Vector Machine for Determining the Health Index of the Insulation System of Inservice Powertransformers[J]. IEEE Transactions on Dielectrics and Electrical Insulation,2013,20(3):965-972.

(编辑:姚树峰)