控制与决策
Control and Decision

基于Fisher Score与最大信息系数的齿轮箱故障特征选择方法

赵玲， 龚加兴， 黄大荣， 胡冲

引用本文:
赵玲, 龚加兴, 黄大荣, 等. 基于Fisher Score与最大信息系数的齿轮箱故障特征选择方法[J]. 控制与决策, 2021, 36(9): 2234–2240.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.1770

基于Fisher Score与最大信息系数的齿轮箱故障特征选择方法

赵 玲[1], 龚加兴[1], 黄大荣[1†], 胡 冲[2]

(1. 重庆交通大学 信息科学与工程学院，重庆 400074；2. 重庆微标科技股份有限公司，重庆 401121)

摘 要: 针对工业环境中齿轮箱多故障特征难以选择的问题,结合Fisher Score与最大信息系数(MIC)构建一种新的故障特征优化选择方法. 首先,考虑到多故障特征分布不均匀和重叠性问题,采用Fisher Score计算方法构建特征指标重要度排序规则;其次,在考虑冗余特征对有效特征表征的影响基础上,利用最大信息系数构建特征间关联性评价方法,对冗余特征实现更新排序;再次,以分类准确率为判断依据,基于支持向量机理论(SVM)对排序模型进行修正,建立基于Fisher Score与最大信息系数的故障特征优化选择方法;最后,利用UCI标准数据集和实验仿真的齿轮箱故障数据进行实验以验证所提出算法的有效性和工程实用性. 仿真实验对比分析表明,与传统的mRMR、reliefF方法相比,所提出的方法特征子集数量适中,准确率更高.

关键词: 齿轮箱；故障特征；Fisher Score；最大信息系数；支持向量机；特征选择

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.1770 **开放科学(资源服务)标识码(OSID):**

引用格式: 赵玲,龚加兴,黄大荣,等. 基于Fisher Score与最大信息系数的齿轮箱故障特征选择方法[J]. 控制与决策, 2021, 36(9): 2234-2240.

Fault feature selection method of gearbox based on Fisher Score and maximum information coefficient

ZHAO Ling[1], *GONG Jia-xing*[1], *HUANG Da-rong*[1†], *HU Chong*[2]

(1. College of Information Science and Engineering，Chongqing Jiaotong University，Chongqing 400074，China；2. Chongqing Micro Standard Technology Co. Ltd，Chongqing 401121，China)

Abstract: Aiming at the problem that it is difficult to select multiple fault features of gearboxes in industrial environment, a new fault feature optimization selection method combining Fisher Score and maximum information coefficient (MIC) is proposed. First, considering about uneven distribution and overlapping of multi-fault features, the Fisher Score calculation method is used to construct the ranking rules of the importance of the feature indicators. Second, based on the impact of redundant features on the effective feature representation, the maximum information coefficient is used to update and rank redundant features. Then, taking classification accuracy as the judgement basis, using the support vector machine (SVM) theory, a fault feature optimization selection method combining Fisher Score and maximum information coefficient is established. Finally, the UCI standard data set and the gear failure simulation data set are used to verify the effectiveness and engineering practicability of the proposed algorithm. Comparative analysis of simulation experiments shows that compared with the traditional mRMR and reliefF methods, the number of feature subsets proposed is moderate and the accuracy is higher.

Keywords: gearbox；fault characteristics；Fisher Score；maximum information coefficient；SVM；feature selection

0 引 言

齿轮传动是机械设备中常用的传动方式,广泛应用于高速列车、风力发电、航空、船舶、石化、矿山、起重运输等行业[1]. 齿轮箱作为机械设备系统的关键部件之一,对其进行精准故障诊断,这对于机械设备正常运行尤为重要. 然而,由于设备自身结构和实际工

收稿日期: 2019-12-19；**修回日期:** 2020-06-09.

基金项目: 国家自然科学基金项目(61703063, 61663008, 61573076)；重庆市技术创新与应用专项重点项目(cstc2019jscx-mbdxX0015)；重庆市教委重点项目(KJZD-K20190070)；重庆市教委科学技术研究项目(KJZD-K201800701)；重庆市研究生科研创新项目(CYS19232)；桥梁工程结构动力学国家重点实验室开放基金项目(2019-01).

责任编委: 胡庆雷.

†通讯作者. E-mail: drhuang@cqjtu.edu.cn.

况的复杂性,齿轮箱故障振动信号往往具有较高的随机性和复杂性,导致故障特征难以准确提取,故障诊断的有效性变得十分困难. 事实上,为提高故障诊断及识别精度,工程上往往通过多角度提取能够反映系统各种状态的多类特征信息来确保故障诊断的最终效果. 但遗憾的是,此类方法不仅导致特征维数增多,在提高相关性的同时也带来特征冗余问题,增加了运算负担,影响识别精度. 显然,在故障数据预处理过程中采用合理方法对多维故障特征进行适当选择,是实际工程故障诊断中的关键环节.

针对特征选择问题,国内外学者做了大量的研究工作,主要集中在以下两类方法: 1) 基于搜索策略划分的特征选择方法; 2) 基于评价准则划分的特征选择方法. 第1类方法通过一种或者两种及以上的基本搜索策略在特征集中搜索特征子集,通常分为全局最优搜索、随机搜索和启发式搜索3种特征选择方式,采用的基本搜索策略有遗传算法、模拟退火、爬山法等[2]. 但在实际操作过程中,由于齿轮箱提取的特征数据复杂,此类特征选择方法存在时间复杂度过高、局部最优的缺陷. 第2类方法主要基于评价准则划分来搜索特征,主要有过滤式(filter)、封装式(wrapper)和嵌入式(embedded)3种方法[3]. 其中: filter方法由于与后续学习算法无关,可直接利用所有训练数据的统计性能评估特征,具有速度快的优势[4-6],但与后续学习算法的性能偏差较大,在大数据特征中故障分类效果不太理想; wrapper方法利用后续学习算法的训练准确率评估特征子集,具备偏差小的优势[7-9],但这类方法计算量大,并不适用大数据集; embedded方法则是将特征选择过程与学习器训练过程融为一体,两者在同一个优化过程中完成,即在学习器训练过程中自动进行特征选择[10],这种特征选择算法在某种程度上兼顾了两类方法的优点,但构造一个合适的函数优化模型是该方法的难点. 此外,主成分分析和线性判别分析等特征变换方法也可以看作一类特殊特征选择方法,但这类方法对特征的理解能力较差.

从上述分析中可以看出,几类方法在特征选择的实际案例中均存在局限性. 为了突破这种局限,本文针对齿轮故障特征选择问题,提出一种基于Fisher Score与最大信息系数(MIC)结合的特征选择方法. 首先,通过Fisher Score评价特征集中所有特征的重要性,并依此进行特征排序; 然后,利用最大信息系数评价特征与特征之间的相关性,从而确定冗余特征,重新对排序结果进行调整; 最后,依据SVM学习算法的分类精度来选择特征子集,并通过实证分析检验本文所提出方法的有效性和合理性. 实验结果表明: 该方法相比于传统的mRMR和reliefF特征选择方法而言,具有一定的优势,改善了齿轮箱故障诊断识别精度不高的问题.

1　相关理论

1.1　Fisher Score相关理论

Fisher Score是一种有效的对样本特征进行评判的标准,传统的Fisher Score源自Fisher线性判别法,其主要思想是在特征集空间中寻找使得不同类别的数据点之间的距离尽可能大,而同一类别之间的距离尽可能小的特征子集[11]. Chen等[12]针对二分类问题,提出了明确的Fisher Score计算方法作为特征选择准则,得到了很好的特征选择效果; 谢娟英等[13]及Güneş[14]分别给出了两种适用于多类问题下的Fisher Score计算方法; Gu等[15]针对特征之间的相关性和冗余性问题,提出了广义Fisher Score; 谢娟英等[16]在所提出多类F-score的基础上,考虑到特征之间的量纲问题,对算法进行了改进; Tao等[17]从类别之间的重叠性和特征的一致性进行考虑,通过引入交叉系数对Fisher Score进行了修正并约简了特征; Song等[18]考虑了多维特征空间中类的相对分布,对Fisher Score进行了改进; 吴迪等[19]考虑到类间分布不均匀和重叠性问题,提出了改进的Fisher Score算法. 本文在其基础上,对其存在的问题进行完善.

针对两类问题,文献[12]提出的Fisher Score值计算方法描述如下. 给定训练样本集$x_k \in R^m, k = 1,2,\ldots,n$,其中$s$类和$q$类的样本数分别为$n_s$和$n_q$. 训练样本第$i$个特征的Fisher Score定义为

$$F_i = \frac{(u_i^s - u_i)^2 + (u_i^q - u_i)^2}{\dfrac{1}{n_s - 1}\displaystyle\sum_{k=1}^{n_s}(x_{k,i}^s - u_i^s)^2 + \dfrac{1}{n_q - 1}\displaystyle\sum_{k=1}^{n_q}(x_{k,i}^q - u_i^q)^2}. \tag{1}$$

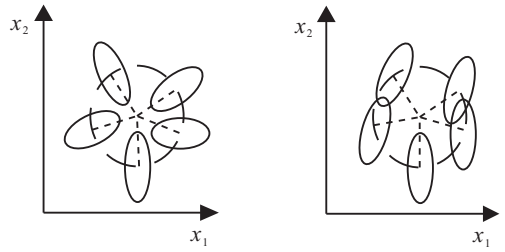
其中: u_i、u_i^s和u_i^q分别是第i个特征在整个数据集上的均值、在s类数据集上的均值和在q类数据集上的均值,$x_{k,i}^s$为第k个s类样本点的第i个特征的特征值,$x_{k,i}^q$为第k个q类样本点的第i个特征的特征值,式(1)的分子为类间散度之和,分母为类内散度之和. Fisher Score值越大,该特征的辨别力越强.

式(1)中的Fisher Score特征选择方法针对两类问题时,未考虑特征在两类之间一致性的问题. 对此,引用文献[17]的交叉系数思想,即

$$N_k = n_{sk} + n_{qk} - n_{sqk}. \tag{2}$$

其中: N_k 表示 s 类与 q 类两类下的特征 x_k 的样本数, n_{sk} 表示 s 类的 x_k 的样本数, n_{qk} 表示 q 类的 x_k 的样本数, n_{sqk} 表示 s 类和 q 类两类特征取值相同的样本数.

在将Fisher Score的计算方法扩展到多类问题时,需考虑类与类之间的分布情况. 针对分布均匀的情况,如图1(a)所示,可使用文献[17]提出的改进的多类Fisher Score的计算方法;针对分布不均匀的情况,如图1(b)所示,通过上述计算方法得到的Fisher Score值与图1(a)的结果是相同的,显然不具合理性.



(a) 分布均匀 (b) 分布不均匀

图 1 多类问题的类间分布情况

文献[18]中提出了一种分布不均匀情况下定义多类之间的类间散度计算方法,即

$$D(x_k) = \sum_{1\leqslant s<q<l} ((n_s + n_q)/N)(u_k^s - u_k^q)^2. \tag{3}$$

其中: $\sum_{1\leqslant s<q<l}$ 代表从类别数 l 中选取两种类别 s 和 q 的所有可能组合,并进行求和; n_s 和 n_q 分别代表第 s 类和第 q 类的样本数; N 为总体样本数; u_k^s 和 u_k^q 分别代表第 k 个特征在第 s 类、q 类样本的均值. 通过式(3)可以更好地表现出类间差异.

综合特征表现为重叠性和分布不均匀的情况,对文献[17]多类Fisher Score值的计算方法进行改写,有

$$F_k = \frac{\sum_{1\leqslant s<q<l} ((n_s + n_q - n_{sq})/N)(u_k^s - u_k^q)^2}{\sum_{j=1}^{l} \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (x_{ki}^j - u_k^i)^2}. \tag{4}$$

其中: N 为去除重复特征值的样本总数,分母为类内散度之和, n_j 为第 j 类的样本数. 修改后的式(4)对特征的重叠性和分布特性进行了考虑,相比于原Fisher Score可以更好地评价特征的重要度. 而文献[19]得出的计算Fisher Score值计算方法如下式所示:

$$F_k = \frac{\sum_{1\leqslant s<q<l} ((n_s + n_q - n_{sq})/N)(u_k^s - u_k^q)^2}{\sum_{j=1}^{l} \sum_{k=1}^{n_j} (x_{kj}^i - u_k^i)^2}. \tag{5}$$

式(5)中虽然对重叠性和分布不均匀的情况进行了考虑,但因协方差与所求类内散度之间存在 $1/(n-1)$ 的系数关系,其中 n 为样本数,故此处直接用协方差方式表达,缺少前置系数,所以导致改进后的Fisher Score作为特征选择评价准则效果并不理想.

1.2 最大信息系数

虽然Fisher Score可以评价特征的重要度,但无法确定特征与特征之间的相关性和特征集中的冗余特征. 现有方法中Pearson系数[20]、最小二乘回归误差和最大信息压缩指数[21]等度量标准被广泛用于度量特征间的线性关系,但都难以刻画特征间大量存在的非线性关系;而信息论中的信息增益[22]、互信息[23]、对称不确定性等度量标准虽然能够同时对特征间线性和非线性关系进行度量,但无法有效度量特征间存在的非函数依赖关系.

Reshef等[24]在2011年提出了一种新的基于信息论的度量标准——最大信息系数. 最大信息系数不仅可以对大量数据中变量间的线性和非线性关系进行度量,还可以广泛地挖掘出特征之间的非函数依赖关系. 最大信息系数方法原理描述如下.

利用互信息和网格划分方法来进行计算. 其中互信息可以看成一个随机变量中包含的关于另一个随机变量的信息量,或者说是一个随机变量由于已知另一个随机变量而减少的不肯定性.

给定两个随机变量 $X = \{x_i, i = 1, 2, \ldots, n\}$ 和 $Y = \{y_i, i = 1, 2, \ldots, n\}$, n 为样本数量,其互信息 $I(X;Y)$ 定义为

$$I(X:Y) = \sum_{x\in X} \sum_{y\in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \tag{6}$$

其中: $p(x,y)$ 为 X 和 Y 的联合概率密度, $p(x)$ 和 $p(y)$ 分别为 X 和 Y 的边缘概率分布密度.

假定一个有限的有序对集合 $D\{(x_i, y_i), i = 1, 2, \ldots, n\}$,将集合 D 中的 x_i 和 y_i 构成的散点图进行 $x \times y$ 的网格划分,分别计算每个网格中的互信息 $I(X:Y)$. 针对 $x \times y$ 的网格划分方式可以取很多种,选取不同划分方式下的 $I(X:Y)$ 的最大值作为划分 $x \times y$ 网格的互信息值. 定义在 $x \times y$ 划分下的网格的最大互信息值记为 $\max(I(X:Y))$;得到最大的互信息值之后,对其进行同时除以 $\log(\min(X,Y))$ 即可完成归一化操作,记为 $\mathrm{mic}(I(X:Y))$,即所求最大信息系数

$$\mathrm{mic}(x,y) = \max_{|X||Y|<B} \frac{\max(I(X,Y))}{\log_2(\min(|X|,|Y|))}. \tag{7}$$

其中: $\max(I(X:Y))$ 表示最大互信息值; B 为网格划分 $x \times y$ 的上限值,是随数据样本数 n 相关的增长函数,文献[24]中取 $B(n) = n^{0.6}$ 效果最好,本文也取该

值.

本文使用最大信息系数来评价特征与特征之间的相关性. 需要给定一个n条样本的特征集$F = \{f_1, f_2, \ldots, f_k\}$,其特征数为$k$. 将特征集类任意两类特征$f_i$和$f_j$的相关性记为$\mathrm{mic}(f_i, f_j)$,$\mathrm{mic}(f_i, f_j)$值越大,说明特征$f_i$与特征$f_j$之间的冗余性越强,可替代性就越强,理想情况下,$\mathrm{mic}(f_i, f_j)$值为0,则说明特征$f_i$与特征$f_j$是相互独立的. 由此,本文对冗余特征的定义如下:

定义1　对于特征集F,若特征f_i与特征f_j的Fisher Score值$F_i > F_j$,且$\mathrm{mic}(f_i, f_j) > 0.8$,则视$f_j$为$f_i$的冗余特征.

2　基于改进的Fisher Score与最大信息系数的特征选择方法

本文提出的基于改进的Fisher Score和最大信息系数的特征选择算法模型如图2所示.

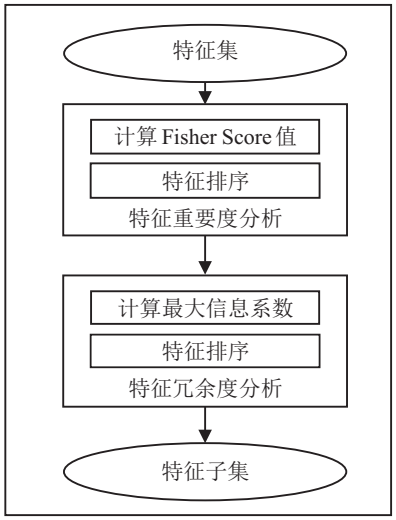


图2　特征选择模型

特征选择算法主要分为两个阶段: 第1阶段,特征重要度分析,按照改进算法分别计算每一个特征的Fisher Score值并按Fisher Score值进行特征排序;第2阶段,利用最大信息系数评价特征与特征之间的冗余性,重新对特征进行排序,按排序后的结果依次选取第1个特征并添加至支持向量机,再使用正向添加策略对特征子集进行扩充,直至添加至最后一个特征,以分类准确率作为特征选择子集的依据. 经过两个阶段选择后的特征子集为最后的特征子集. 基于改进Fisher Score与最大信息系数的特征选择方法可描述如下:

输入: 特征数据集$F(x_1, x_2, \ldots, x_k)$;

输出: 最优故障特征子集F_{out}.

第1阶段: 特征重要度分析.

step 1: 通过式(4)依次计算特征数据集F中对应的Fisher Score值F_k;

step 2: 对特征集F按F_k进行降序排列.

第2阶段: 特征冗余度分析.

step 1: 按照F排序,对F进行遍历,逐次选取F_k值较大的特征f_i和比其F_k值小的特征f_j,根据式(7)计算$\mathrm{mic}(f_i, f_j)$;

step 2: 判断是否$\mathrm{mic}(f_i, f_j) > 0.8$,如果大于0.8,则将特征$f_j$顺序$F'$调整至末端,更新$F$并排序,同时记录$F_j$不参与$\mathrm{mic}(f_i, f_j)$值的计算;

step 3: 遍历完成,输出F_{out}.

算法流程如图3所示.

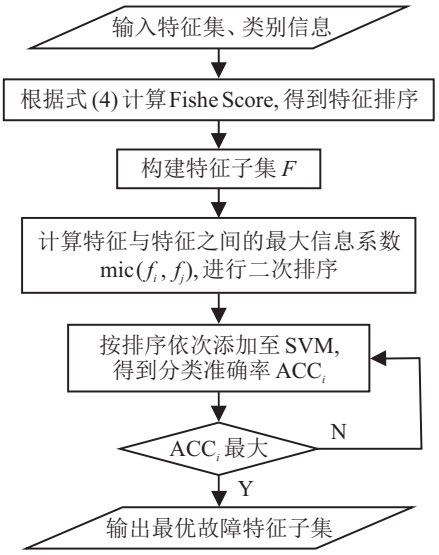


图3　特征选择框架

与其他特征选择方法相比,本文提出的特征选择方法充分利用了Fisher Score计算简单的优势,改进后普适性更高;考虑到特征冗余性问题,引入最大信息系数,相比于其他方法,在复杂关系的数据集上的相关性评价更为准确,在复杂数据集上的表现从理论上效果更好.

3　仿真及实例验证

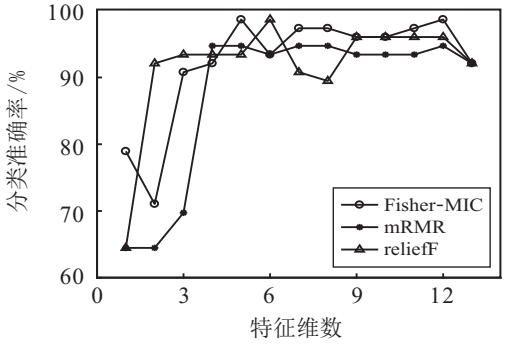
本文使用UCI机器学习数据库[25]和自建齿轮故障特征数据集进行实验验证,前者是特征选择工作中广泛使用的实验数据,作为本文进行实验验证的标准数据. 特征选择方法通常使用分类器的准确率来评价所选取的特征子集的好坏,因此,本文在实验中采用SVM分类器进行分类和评价,利用分类准确率作为所选择的特征子集的优劣评价标准,分类准确率由交叉验证获得.

3.1　仿真验证

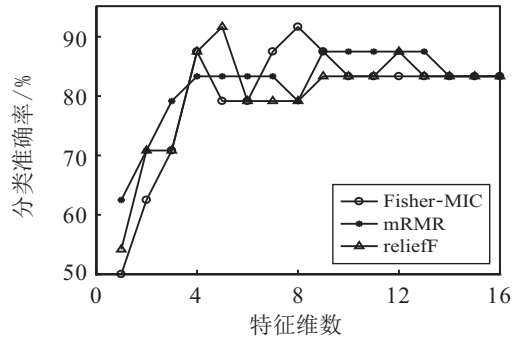
为验证基于改进的Fisher Score与最大信息系数的特征选择方法的有效性,选择最大相关最小冗余(mRMR)算法和reliefF算法进行实验对比,利用UCI

机器学习数据库3个常用的数据集Wine、 Zoo和Ionosphere作为分析对象. 其中: Wine数据集样本数为178, 类别数为3, 特征数为13; Zoo数据集样本数为101, 类别数为7, 特征数为16; Ionosphere数据集样本数为351, 类别数为2, 特征数为33. 数据集包含了二分类和多分类数据, 样本数最低为101, 最高为351, 可从多方面评估特征选择方法的有效性.

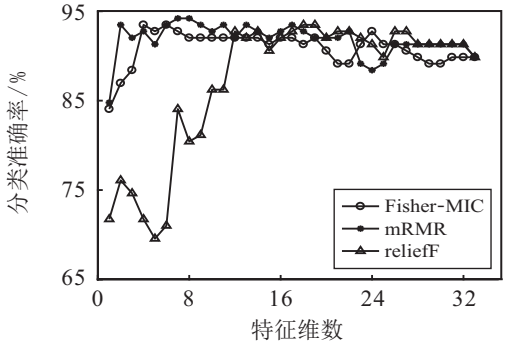
采用随机抽取的方式将整个数据集划分为成6∶4的训练数据集和测试数据集, 使用本文提出的特征选择方法(Fisher-MIC)和mRMR算法以及reliefF算法从训练集中选出特征子集, 并将选出的特征子集应用到测试集进行测试. 实验中使用5折交叉验证对特征子集进行测试和评价. 为避免选出的数据出现重复选取问题, 重复上述实验过程10次, 对10次的实验结果求均值得到最后的实验结果. 不同特征选择方法在特征数据集的特征选择效果如图4所示.



(a) Wine数据集



(b) Zoo数据集



(c) Ionosphere数据集

图4 数据集测试结果

如图4所示, Fisher-MIC方法在3个数据集上表现略有不同: 在Wine数据集上表现出的分类准确率最高, 且选择的特征集最少; 在Zoo数据集上与采用reliefF方法达到的准确率相当, 但选择的特征维数略高; Ionosphere数据集上则是mRMR方法表现效果较好. 究其原因: Zoo数据集因其实际类别为不同动物, 数据中的101类动物特征表现上不具有重叠性, 故Fisher-MIC相比reliefF而言, 虽然最后选择出的特征子集的分类效果与reliefF相同, 但所选特征维数略高; 在Ionosphere数据集上, 3种特征选择方法最终选择的特征子集表现相差不大, 但在特征个数选取较少的情况下, Fisher-MIC得到的特征子集具有后续计算量较小、分类较好的优势, 特征选择结果如表1所示.

表1 UCI数据集测试结果 %

数据集	指标	所有特征	mRMR	reliefF	Fisher-MIC
Wine	accuracy	92.11	94.73	98.68	98.68
	特征数	13	4	6	5
Zoo	accuracy	83.33	87.55	91.667	91.667
	特征数	17	9	5	8
Ionoshphere	accuracy	89.85	94.42	94.42	93.47
	特征数	34	7	18	4

表1中的结果显示, 本文提出的特征选择方法能对标准数据集(Wine、Zoo、Ionoshphere)的数据进行有效的特征选择, 经特征选择之后特征维数大大减少, 而且分类精度也有一定的提升, 在部分数据集上表现效果相比其他方法而言具有计算量较低的优势. 从仿真结果看, 本文提出的特征选择方法对齿轮故障特征进行特征选择在理论上是可行的.

3.2 实例验证

为进一步验证Fisher Score结合最大信息系数的特征选择方法的有效性, 利用故障模拟平台上测取的数据对所提出的方法进行验证. 信号主要包括齿轮齿面点蚀、齿面磨损、齿面断齿等故障信息.

基于以上故障种类, 通过图5仿真实验平台采集故障信号, 振动信号采样频率 $f_s = 5\,210$ Hz, 齿轮转速为 $r = 880\,\mathrm{r/min}$, 大齿轮的模数 $m = 2$, 齿数

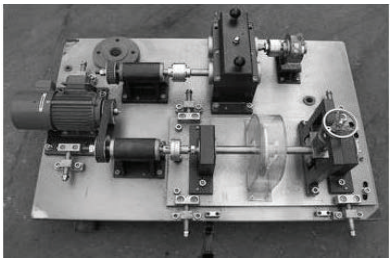


图5 仿真实验平台

$c_n = 75$;小齿轮模数$m = 2$,齿数$c_n = 55$. 振动信号包括:大齿轮正常信号、大齿轮点蚀故障信号、大齿轮断齿故障信号、小齿轮磨损故障信号.

仿真实验的特征指标通常通过时域分析方法和频域分析方法获得,在区分故障特征时,单一的时域特征和频域特征对大齿轮点蚀故障信号以及小齿轮磨损故障信号难以区分,所以引入去趋势多重分形对故障信号进行特征提取. 从时域分析方法、频域分析方法以及多重分形等3个角度对障信号进行特征提取. 实验选取120个样本,分别选取了4类齿轮运行状态:正常运行状态、磨损运行状态、点蚀运行状态和断齿运行状态,每类状态各30组样本. 选取齿轮不同状态下的时域特征16个(无量纲指标:波形指标、峰值指标、裕度指标、脉冲指标和歪度指标;有量纲指标包括最大值、平均值、均方值、均方根值、方差、最小值、峰值、均方幅值、方根幅值、平均幅值、峭度)、频域特征5个(均方根频率、标准差频率、重心频率、均方频率和方差频率)、多重分形谱特征9个(奇异指数、最大奇异指数、最小奇异指数、谱宽、谱最大值、最大最小概率多重分形谱的差、广义Hurt参数$h(q)$、$\sum h(q)$及$h(\bar{q})$)共计30维特征作为输入特征.

对上述特征数据分别采用Fisher-MIC、mRMR和reliefF特征选择方法从训练集中选择出特征子集,并将选择出的特征子集应用到测试集进行测试,最终特征选择结果如表2所示.

表2　齿轮故障特征集测试结果　　%

数据集		所有特征	mRMR	reliefF	Fisher-MIC
故障特征集	accuracy	79.17	87.50	83.33	91.667
	特征数	30	23	5	8

由表2数据,对比特征选择前后的分类精度可知,未选择齿轮数据特征时,选择前的特征准确率为79.17%;分别采用mRMR、reliefF及Fisher-MIC特征选择方法进行特征选择后再进行分类的正确率均有所提高,相比于mRMR和reliefF方法而言,本文方法的特征选择个数适中且选择出的特征子集分类效果更好.

4　结　论

1) 对Fisher Score计算方法改进后,作为新的特征评价准则,计算量依然较小且更具有普适性.

2) 选择最大信息系数评价特征相关性,能有效地评价特征与特征之间的关系和确定冗余特征.

3) 在结论1和结论2的基础上,利用支持向量机对特征子集进行评价,选出最优特征子集,兼顾了filter和wrapper的优点,与传统mRMR和reliefF方法对比,具有准确率高、计算量小的优势.

虽然本文方法得到了一些进步,但仅初步应用于齿轮箱故障数据集,在其他一些数据集上测试,仍存在一定问题;最大信息系数评价特征之间的冗余性的阈值确定,仍需根据实际情况考虑. 本文作者目前正在做这方面的研究,限于篇幅,将另文给出.

参考文献(References)

[1] 雷亚国, 何正嘉, 林京, 等. 行星齿轮箱故障诊断技术的研究进展[J]. 机械工程学报, 2011, 47(19): 59-67.
(Lei Y G, He Z J, Lin J, et al. Research advances of fault diagnosis technique for planetary gearboxes[J]. Chinese Journal of Mechanical Engineering, 2011, 47(19): 59-67.)

[2] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述[J]. 控制与决策, 2012, 27(2): 161-166.
(Yao X, Wang X D, Zhang Y X, et al. Summary of feature selection algorithms[J]. Control and Decision, 2012, 27(2): 161-166.)

[3] Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution[C]. Proceedings of the 20th International Conference on Machine Learning. Washington DC, 2003: 856-863.

[4] Osanaiye O, Cai H B, Choo K K R, et al. Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing[J]. EURASIP Journal on Wireless Communications and Networking, 2016(1): 1-10.

[5] Ambusaidi M A, He X J, Nanda P, et al. Building an intrusion detection system using a filter-based feature selection algorithm[J]. IEEE Transactions on Computers, 2016, 65(10): 2986-2998.

[6] 张俐, 王枞. 基于最大相关最小冗余联合互信息的多标签特征选择算法[J]. 通信学报, 2018, 39(5): 111-122.
(Zhang L, Wang Z. Multi-label feature selection algorithm based on joint mutual information of max-relevance and min-redundancy[J]. Journal on Communications, 2018, 39(5): 111-122.)

[7] Mafarja M, Mirjalili S. Whale optimization approaches for wrapper feature selection[J]. Applied Soft Computing, 2018, 62: 441-453.

[8] Hu Z Y, Bao Y K, Xiong T, et al. Hybrid filter−wrapper feature selection for short-term load forecasting[J]. Engineering Applications of Artificial Intelligence, 2015, 40: 17-27.

[9] 杨宇, 潘海洋, 程军圣. 基于特征选择和RRVPMCD

的滚动轴承故障诊断方法[J]. 振动工程学报, 2014, 27(4): 629-636.
(Yang Y, Pan H Y, Cheng J S. The rolling bearing fault diagnosis method based on the feature selection and RRVPMCD[J]. Journal of Vibration Engineering, 2014, 27(4): 629-636.)

[10] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 252-254.
(Zhou Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 252-254.)

[11] Tsuda K, Kawanabe M, Müller K R. Clustering with the fisher score[C]. Advances in Neural Information Processing Systems. Vancouver, 2003: 745-752.

[12] Chen Y W, Lin C J. Combining SVMs with various feature selection strategies[C]. Feature Extraction. Berlin, Heidelberg: Springer, 2006: 315-324.

[13] 谢娟英, 王春霞, 蒋帅, 等. 基于改进的F-score与支持向量机的特征选择方法[J]. 计算机应用, 2010, 30(4): 993-996.
(Xie J Y, Wang C X, Jiang S, et al. Feature selection method combing improved F-score and support vector machine[J]. Journal of Computer Applications, 2010, 30(4): 993-996.)

[14] Güneş S, Polat K, Yosunkaya Ş. Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome[J]. Expert Systems with Applications, 2010, 37(2): 998-1004.

[15] Gu Q Q, Li Z H, Han J W. Generalized fisher score for feature selection[J]. 2012: arXiv:1202.3725.

[16] 谢娟英, 雷金虎, 谢维信, 等. 基于D-score与支持向量机的混合特征选择方法[J]. 计算机应用, 2011, 31(12): 3292-3296.
(Xie J Y, Lei J H, Xie W X, et al. Hybrid feature selection methods based on D-score and support vector machine[J]. Journal of Computer Applications, 2011, 31(12): 3292-3296.)

[17] Tao P, Yi H, Wei C, et al. A method based on weighted F-score and SVM for feature selection[C]. The 25th Chinese Control and Decision Conference (CCDC). Guiyang: IEEE, 2013: 4287-4290.

[18] Song Q J, Jiang H Y, Liu J. Feature selection based on FDA and F-score for multi-class classification[J]. Expert Systems with Applications, 2017, 81: 22-27.

[19] 吴迪, 郭嗣琮. 改进的Fisher Score特征选择方法及其应用[J]. 辽宁工程技术大学学报: 自然科学版, 2019, 38(5): 472-479.
(Wu D, Guo S Z. An improved Fisher Score feature selection method and its application[J]. Journal of Liaoning Technical University: Natural Science, 2019, 38(5): 472-479.)

[20] Saqlain S M, Sher M, Shah F A, et al. Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines[J]. Knowledge and Information Systems, 2019, 58(1): 139-167.

[21] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF[C]. European Conference on Machine Learning. Berlin, Heidelberg: Springer, 1994: 171-182.

[22] Mitra P, Murthy C A, Pal S K. Unsupervised feature selection using feature similarity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 301-312.

[23] Yang Y M, Pedersen J O. A comparative study on feature selection in text categorization[C]. Proceedings of the 14th International Conference on Machine Learning. Nashville, 1997: 412-420.

[24] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets[J]. Science, 2011, 334(6062): 1518-1524.

[25] Lichman M. UCI machine learning repository[EB/OL]. [2019-12-09]. http://archive.ics.uci.edu/ml.html.

作者简介

赵玲 (1979−), 女, 副教授, 博士, 从事动态系统的故障诊断与容错控制、复杂系统的分析与设计等研究, E-mail: zhao.ling@163.com;

龚加兴 (1996−), 男, 硕士生, 从事动态系统的故障诊断与容错控制的研究, E-mail: gjx18084061690@gmail.com;

黄大荣 (1978−), 男, 教授, 博士生导师, 从事动态系统的故障诊断与容错控制、复杂系统的分析与设计等研究, E-mail: drhuang@cqjtu.edu.cn;

胡冲 (1983−), 男, 高级工程师, 硕士, 从事动态系统的故障诊断与容错控制的研究, E-mail: huchong@cqrfid.cn.

(责任编辑：李君玲)