DOI: 10.3901/JME.2019.17.133

面向制造过程数据的两阶段无监督特征选择方法*

张洁^{1,2} 盛夏¹ 张朋¹ 秦威¹ 赵新明¹

(1. 上海交通大学机械与动力工程学院 上海 200240;

2. 东华大学机械工程学院 上海 201620)

摘要: 现代化制造车间无时无刻不在产生大量数据, 其中绝大部分以无标签结构化原始数据的形式存储在现代化制造企业的工业大数据平台中。这些制造数据一方面具有很大的潜在价值, 另一方面因为其具有高噪声、高冗余性的特点, 难以直接分析与利用。因此, 针对制造过程原始数据的特点, 以去除制造数据冗余性、挖掘原始数据局部结构为目的, 提出一种两阶段无监督特征选择方法。该方法的第一阶段采用遗传算法产生的原始数据的低维子集作为径向基神经网络(Radial basis function neural network, RBFNN)的输入, 利用 RBFNN 复现原始数据的全部维度, 并以降维率及复现精度作为遗传算法(Genetic algorithm, GA)的适应度函数, 通过 GA 多次迭代学习高维特征的低维表示, 删除原始数据集中的冗余特征与噪声特征。第二阶段采用拉普拉斯特征得分(Laplacian score, LS)逐维评价剩余特征对于反映数据局部几何结构的作用, 挖掘对改善分类性能更有效的特征。通过与 LS 等无监督特征选择算法对比, 验证了提出的两阶段无监督特征选择方法能够有效降低制造数据的冗余性, 并提高数据的分类性能。

关键词: 无监督特征选择; 遗传算法; 径向基神经网络; 拉普拉斯得分; 制造过程数据

中图分类号: TG156

Two-stage Unsupervised Feature Selection Method Oriented to Manufacturing Procedural Data

ZHANG Jie^{1,2} SHENG Xia¹ ZHANG Peng¹ QIN Wei¹ ZHAO Xinming¹

(1. Institute of Intelligent Manufacturing and Information Engineering,

Shanghai Jiao Tong University, Shanghai 200240;

2. College of Mechanical Engineering, Donghua University, Shanghai 201620)

Abstract: In a modernized manufacturing workshop, myriads of data are incessantly produced and a large part of those are stored in the industrial big data platform of the modern manufacturing enterprise in the form of structuralized unlabeled raw data. Those manufacturing data are of great latent exploitative value, because of their characteristics of high-noise and high-redundancy, however, direct analysis and utilization of them are impossible. Aiming at reducing the redundancy of manufacturing procedural data and excavating their local structure, a two-stage unsupervised feature selection method is proposed. In the first stage of the method, subset of the original feature set generated by genetic algorithm(GA) is utilized as the input features of radius basis function neural network(RBFNN), to reconstruct the unabridged original feature set. The ratio of dimensionality reduction and reconstruction accuracy are calculated jointly as the fitness function of GA, which is optimized by iteration to learn a low-dimensional representation of high-dimensional features, removing redundant and noisy features of the origin feature set. In the second stage, Laplacian score(LS) is employed to evaluate the locality preserving power of the remaining features, unearthing features which are prone to improving the performance of classification. By comparing with other unsupervised feature selection method, the method proposed here is proven more effective in reducing the redundancy of manufacturing data and simultaneously enhancing the performance of classification.

Key words: unsupervised feature selection; genetic algorithm; radius basis function neural network; Laplacian score; manufacturing procedural data

* 国家自然科学基金资助项目(U1537110, 51435009)。20180919 收到初稿,
20190203 收到修改稿

0 前言

随着计算机技术的不断发展及其与制造业的深度融合,我国的现代化制造车间逐步完成了制造业信息化建设,实现了不同生产阶段数据的采集、处理、存储与应用^[1]。这些数据经过传感器、RFID 等异源数据采集设备采集后,经过数据集成与简单预处理,以原始数据的形式存储在制造企业的工业大数据平台中,并成为故障诊断^[2]、质量控制^[3]等高层次制造数据挖掘应用的数据来源。制造过程的原始数据来源广泛且重复性极高,同一制造特性往往由不同数据采集设备重复多次采集,产生极大的冗余性,降低了数据的价值密度。尽管数据集成过程已经减少了一部分由于数据源的异源性而产生的语义重复与语义错误,但在大数据平台中存储的原始数据依旧不可避免地呈现海量、高维、高冗余性、低价值密度的制造业大数据^[1]特点,难以直接利用与挖掘。另外,相互冗余的制造特性之间,由于异源传感器的误差分布不同,其冗余性潜藏在沿同质传感器分布的噪声之下。这些冗余特性在数据集成阶段很难得到消除,进一步提升了识别制造数据间的冗余性、删除冗余特征的难度。

此外,在高层次数据应用目标尚未明确的情况下,为了避免数据丢失对今后的业务应用的影响,企业通常选择保留全部历史数据。然而数据的价值是由数据的潜在结构决定的,由于制造原始数据中天生存在着大量变化不大的无关特征,保留这些特征对于最终的业务应用无太大作用,增加额外存储空间,甚至由于无关维度的增多,引发高维数据的“维度灾难”^[4],降低了某些对维度敏感的大数据分析算法的精度,阻碍了大数据分析与应用业务的开展。通过对制造数据特征的筛选,有效降低制造过程原始数据的冗余性与无关性,对于减少制造数据存储空间、提高制造数据价值密度及提升数据应用业务效果具有重要意义。

制造特征的筛选问题属于特征选择问题。根据数据是否具有类别信息等输出标签^[5],可将特征选择分为有监督特征选择与无监督特征选择。在制造业中应用的特征选择方法以有监督特征选择方法为主,并被广泛应用于故障诊断^[2]、质量控制^[3]、工期预测^[6-7]等工程问题中。针对不同的业务场景,不同研究设计了不同的特征选择算法。针对故障诊断问题实时性的要求,ISLAM 等^[8]设计了并行化判别式特征选则方法,减少了计算时间,并在实时故障诊断场景中证明了该算法的有效性。闫伟^[3]针对质量

控制中的数据不平衡问题,提出了关键质量特性识别(Critical to quality, CTQ)特征选择方法,成功地降低了将次品误判为合格品的第二类错误发生的概率。WANG 等^[6]针对海量、高维大数据场景,设计了轻量化的特征选择方法,在实现精确工期预测的同时,降低了算法的复杂度。由上述文献可知,有监督特征选择算法尽管针对性强,但面对不同的业务场景需要设计不同的特征选择方法,泛化性不强。此外,为数据加标签的过程还需依赖专家评定,耗时长且成本较高。制造过程数据绝大部分是无标签数据。相比有监督特征选择算法,无监督特征选择^[9-10,12]不需要事先对数据做标签,成本更低,泛化性更强,且能挖掘数据的潜在特征,对于提高数据理解性、解决高维灾难和提升分类效果更有意义。

传统的无监督特征选择算法如拉普拉斯得分法独立地评价每一个特征维度,并根据评价指标的得分高低筛选特征。当特征中存在冗余特征时,相互冗余的特征会得到相近的得分,使得最后筛选出的特征都是具有相似保留样本局部特性能力的冗余特征。针对这一问题,近些年的文献[10, 13-14]采用多任务学习的形式联合评估多个特征维度的组合特征对响应矩阵 S 的表示能力,并通过 L_1 范数为表示系数矩阵施加稀疏性约束,从而使得所选特征子集在保留样本局部特性的前提下,冗余性下降。通过选用不同的响应矩阵 S 及不同的范数约束,可以达到不同的选择目的。ZHU 等^[9]提出的正则自表示特征选择方法取响应矩阵为原始数据集,设计优化目标为 $L_{2,1}$ -范数,利用 L_1 -范数的稀疏性效果限制每次表达中选择的特征个数,利用 L_2 -范数的正则化效果限制表达特征权重的大小,通过计算特征评分去除冗余特征和对自表示更有效的特征。ZHAO 等^[10]通过设计响应矩阵为样本相似度矩阵,结合 $L_{2,1}$ -范数,在同一个模型中同时达到去冗余性和保留样本局部特性的目的。这些无监督特征选择算法很好的解决了线性冗余的问题,但针对不同的响应矩阵与不同的范数约束,需要设计不同的优化算法,且由于优化目标的非凸性,算法较为复杂。此外,这些算法不能降低非线性冗余性。考虑到实际的制造过程数据以非线性冗余为主,传统的无监督特征选择方法不能满足实际需要。

因此,针对传统特征选择算法的不足,结合制造过程数据无标签、高冗余、高噪声的特点,提出两阶段无监督特征选择算法,对制造过程中的原始数据进行降维处理,使得降维后的数据集具有较低的冗余性及较高的相关性,为更高层次的数据挖掘应用提供数据基础。

1 问题描述

1.1 制造过程数据特点分析

制造过程数据^[1]是指在车间加工、装配等过程中,由生产活动产生的机器状态、加工状况等数据,这些数据由机床传感器测得或由工人记录,并以原始数据的形式存储在MES、ERP系统中。这些数据的产生方式不同,各自反应了制造过程中的某一方面特性,常常作为制造质量控制、故障诊断等数据挖掘应用的数据来源。

制造过程数据具有无标签、高冗余、高噪声的特点。绝大部分的制造数据是不具有目的性的原始数据,不存在输出标签。即使是通过专家对数据类别进行标记,由于标定类别的成本较高、周期长,通常只会对选择对其中一小部分样本进行标记。因此,为了有效利用占绝大部分的无标签制造过程数据,必须采用无监督特征选择方法。

制造过程数据还包含了较多的冗余信息,其冗余性主要来源于多传感器的重复采集、异源数据库数据归并或可推导特征的重复收集。多源传感器的重复采集如不同机器对于加工产品ID的表述形式不同,但本质并无区别。异源数据库的归并发生在数据集成阶段,多个数据库内的数据按主键发生归并时可能会引入重复字段;可推导特征例如发动机数据集中,同时统计了“额定功率”、“额定转速”、“额定扭矩”,其中的一个维度完全可以由其他维度非线性表出,因此同时采集这三个特征就会引入冗余性。冗余信息会增加计算时间、降低算法精度,必须要进行去除。

此外,由于传感器之间的误差分布不同,可能造成某两个特征之间虽然相互冗余,但其中一维存在噪声,导致无法识别该冗余性的问题。设计的特征选择算法应当能够识别这种隐藏在沿特征分布的误差之下的冗余性。

因此,针对制造过程数据的这三个重要特点,采用无监督特征选择方法对无标签数据进行特征筛选,识别噪声特征,去除冗余特征,挖掘有价值特征。下小节着重介绍无监督特征选择问题的定义、目标、输入与输出。

1.2 无监督特征选择问题

无监督特征选择问题^[9-10]的定义为,对于给定无标签数据集的全部特征维度,选择其中的理想特征子集,使这一子集具有最小冗余性和最大相关性。对于无监督特征选择,冗余性指的是特征间的相关性^[11],包括线性相关性与非线性相关性。相关性^[12]

则是指单个特征对于反映数据样本局部结构作用的大小。无监督特征选择的数学模型如下所述。

设输入为归一化后的结构化数据集 $\mathbf{X} \in \mathbf{R}^{n \times m}$,其中 n 代表数据集的样本个数, m 代表数据集的特征维度。 $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ 代表 n 个样本,则 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T$ 。特征集合 $\mathcal{F} = \{f_1, f_2, \dots, f_m\}$ 代表 $\mathbf{X}$ 的 m 维特征。 x_{ij} 为第 i 个样本在第 j 个特征上的取值, $x_{ij} \in [0, 1]$ 。

数据集 $\mathbf{X}$ 特征的冗余性定义为:若对于特征 f_i ,存在非线性函数 Φ 及全集 $\mathcal{F}$ 的子集 $\mathcal{S} = \{f_1, f_2, \dots, f_k\} (k < m)$,使得 $\Phi(f_1, f_2, \dots, f_k) = f_i$,则 f_i 对于全集 $\mathcal{F}$ 是冗余的。

数据集 $\mathbf{X}$ 特征的相关性定义为该特征保留数据局部结构的能力,即保留样本相似性的能力⁰。设 $\mathbf{S}$ 为基于某种相似度定义的样本相似度矩阵,其最大特征值对应的特征向量包含了最多的样本分部信息,因此与该矢量方向近似的特征具有最强的保留样本相似性的能力,即具有最强相关性。在具有较强相关性的维度上,数据呈现明显的多簇分布特性,即簇内点距离小,簇间点距离大。

由于冗余性与相关性是不冲突的两个性质,不妨采用两阶段的特征选择方法。第一阶段主要消除原始数据集中的非线性冗余特征,第二阶段对特征逐个评价,筛选相关特征,既能有效避免冗余性对于相关性筛选的影响,又能避免在第二阶段使用复杂的优化算法。因此,下一部分提出两阶段无监督特征选择方法实现这一思路,使得最终的输出为具有最小冗余性与最大相关性的特征子集 $\mathcal{S}$ 。

2 制造数据两阶段特征选择方法

图1为制造数据两阶段特征选择方法(TSFS)的整体框图。特征选择方法的输入是制造数据经过简单预处理(如去重复值、无效值等)后得到的原始特征集,包含独立特征(如图1中 f_1, f_2, f_3, f_4)、组合冗余特征(如图1中 $F(f_2, f_3)$)及含噪声的冗余特征(如图1中 $\tilde{f}_1$)。方法的第一阶段通过遗传算法与径向基神经网络结合去除原始数据集中的冗余特征(如 $F(f_2, f_3), \tilde{f}_1$),获得无冗余特征集(如最理想情况 $\{f_1, f_2, f_3, f_4\}$),第二阶段对无冗余特征集中的特征逐个进行打分,评价其保留样本相似性的能力。由于算法本身较为抽象,为了便于与实际业务场景相结合,首先给出由实际制造数据集简化后的案例,便于明确算法输入、算法各阶段目的及算法的理想输出。之后从理论上分两个部分深入介绍本文提出的两阶段特征选择方法。

图 1 两阶段无监督特征选择算法框架图

2.1 算法实际输入案例及理想输出

表 1 为某柴油发动机制造厂大数据平台数据仓库中节选的数据表, 该表中所统计的制造特征主要来自发动机质量产品检验环节人为测量或监测设备的测量值。

表 1 柴油发动机质量检测数据表

发动机 编号	缸套突出 高度/mm	机体泄漏 量/L	标况大气 压/MPa	标况转速 /(r/min)	标况功率 /kW	标况扭矩 /(N·m)
L01	0.129	5.47	10.06	2 197	228.4	993
L02	0.127	10.98	10.06	2 200	226.2	982
L03	0.124	2.52	10.06	2 197	225.7	981
L04	0.131	8.13	10.06	2 198	224.8	977
L05	0.120	13.13	10.06	2 196	223.6	972
L06	0.129	1.47	10.07	2 199	223.7	973
L07	0.131	20.82	10.06	2 198	223.2	971
L08	0.117	96.50	10.07	2 202	223.4	970
L09	0.134	3.16	10.06	2 198	222.1	969
L10	0.126	5.78	10.07	2 198	222.1	965
L11	0.129	10.77	10.06	2 198	221.8	965
L12	0.153	2.35	10.06	2 197	222	964

由表 1 可知, 缸套露出高度、泄露量、标况大气压为独立特征, 标况功率、标况转速及标况扭矩可以由其中的两个特征推导出另一个特征, 因此这三个特征中有一个特征是冗余的。由于角速度传感器与力传感器误差模型不同, 因此利用理论公式由三个特征中的两个特征推导另一个特征会发现存在一定的偏差, 但三者内部的冗余性是不变的, 只是这种冗余性被噪声掩盖了。此外标况大气压的所有样本值变化不大, 且分布稀疏, 因此该维度相较于其他维度对于最终应用(如良率计算)并无太大作用。

设表 1 的七个特征分别为 $f_1 \sim f_7$ 。由上面分析可知, 显然 f_6 满足 $f_6 = F(f_5, f_7) + \varepsilon$, 那么本文算法的输入就是全部特征 $f_1 \sim f_7$ 。算法一阶段去除数据集中的冗余性, 那么算法理想输出应是除去标况转速 f_5 (或标况功率 f_6 , 或标况扭矩 f_7) 的剩余其他特征维度, 即标况转速(或其他情况)作为冗余特征被删除。算法的第二阶段去除特征中的无用特征, 即理想输出结果应是保留其他特征($f_1 \sim f_3$ 、 f_6 、 f_7)的特征, 标况大气压 f_4 被删除。

2.2 基于 GA-RBFNN 的第一阶段特征选择方法

典型的特征选择方法^[15]一般包括子集生成、特征评价、终止判断和结果验证四个步骤, 其中前两个步骤是特征选择方法设计的关键。

子集生成是指采用一定的搜索策略产生候选特征子集的过程, 常用的生成方法有全局最优搜索、随机搜索、启发式搜索等。以遗传算法为代表的随机搜索方法较全局搜索方法的搜索空间更少, 并且与启发式方法相比能够选出更优秀的特征子集, 被学者们广泛采用^[16-18]。特征评价方法则以过滤法^[19]为主, 通常采用距离度量^[9]、信息度量^[20]、依赖性度量等。

本文设计的第一阶段特征选择方法采用遗传算法作为子集产生方法产生候选特征作为 RBF 神经网络的输入, 复现全部特征, 利用其复现精度及候选特征的降维率之和作为适应度函数筛选优秀个体。经过一定世代的迭代, 将最终世代的最优个体作为特征选择问题的最终解。一阶段算法的原理图如图 1 所示左半部分所示。第 2.2.1 节介绍 RBF 神经网络的原理, 第 2.2.2 节介绍 GA-RBFNN 特征选择算法设计, 第 2.2.3 节介绍 RBF 神经网络的缺陷

及特征选择模型解决高维问题时的变化。

2.2.1 径向基神经网络

神经网络是一种拟合复杂非线性关系的常用工具, 具有较强的泛化性。当输出能够被输入完全非线性表出时, 神经网络具有较高的拟合精度。径向基神经网络是一种三层前馈神经网络, 包含输入层、隐藏层和输出层, 其结构如图2所示。

图2 径向基神经网络结构

其中输入层到隐藏层是无权全连接, 直接将输入特征输入隐藏层神经元。隐藏层中, 第 i 个神经元的激活函数 $\Phi_i(\mathbf{x})$ 是高斯核函数, 其表达式如下

$$\Phi_i(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_i\|^2}{\sigma^2}\right) \quad (2)$$

其中, $\mathbf{x}_i \in \mathbf{R}^m$ 是维的输入样本矢量, $\mathbf{c}_i$ 是第 i 个神经元径向基函数中心, 可直接去样本矢量或采用聚类法求得, σ 是拓展常数。由于高斯核函数是无限维非线性函数, 因此只要满足 $\forall i \neq j, \mathbf{c}_i \neq \mathbf{c}_j$, 则任何一个径向基函数都可以作为无限维空间中的一个基, 从而对于这组基进行线性加权求和可以逼近任意一个函数。网络输出层的输出值由式(3)计算

$$\hat{\mathbf{y}}(k) = \sum_{i=1}^L w_{ki} \Phi_i(\mathbf{x}) \quad (3)$$

其中, $k=1, 2, \dots, K$ 代表输出层节点个数, w_{ki} 代表第 i 个隐藏层节点与第 k 个输出节点之间的连接去权值。式(3)的矩阵形式如式(4)

$$\mathbf{Y} = \mathbf{W}^T \Phi(\mathbf{X}) \quad (4)$$

当样本个数大于权值未知数个数时, 上述方程的最小二乘解为

$$\mathbf{W} = \Phi(\mathbf{X})^+ \mathbf{Y}^T \quad (5)$$

由式(5)得到的解是权值矩阵 $\mathbf{W}$ 的全局最优解, 因此与BPNN等神经网络相比具有更快的训练速度, 且不存在陷入局部最优的问题。

2.2.2 GA-RBFNN 算法设计

遗传算法^[22]是一种通过模拟生物进化过程自然选择和遗传机理寻找最优解的方法, 通常分为基因编码、初始化种群、适应度计算、选择、交叉、变异六个步骤。它首先随机产生一组初始解, 统称为种群。种群中的每一个解称为个体, 每一个个体通过某种编码方式编成染色体。评价解的优良程度函数称为适应性函数。通过计算每个初始解的适应性函数, 高概率选择适应度高的个体, 并在选择的个体染色体之间交叉、变异产生下一代个体, 完成一代进化过程。遗传算法的收敛条件是达到一定的世代数或适应性函数达到一定阈值。

遗传算法的个体基因编码采用二进制编码的形式。若选择特征 f_i , 则染色体位置第 i 个基因置1, 否则置0。图3代表以7个制造特征作为问题输入时, 选中第一至第三及第六个特征。

1	1	1	0	0	1	0
---	---	---	---	---	---	---

图3 遗传算法染色体编码案例

特征选择算法的目标包括获得较高的RBFNN复现精度, 同时保持较高的特征降维率, 因此GA的适应度函数应当是这两个优化目标的折中。在文献[23]提出的最终预测误差(Final prediction error, FPE)函数的基础上, 设计了新的适应度函数, 如式(6)~(8)

$$F(pop) = R(pop) + M(pop) \quad (6)$$

$$R(pop) = \text{sigmoid}\left(\frac{\alpha(m - 2m')}{m}\right) \quad (7)$$

$$M(pop) = \text{sigmoid}\left(\frac{-\lg\left(\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2\right) - d}{s}\right) \quad (8)$$

式中, pop 为种群中的个体, $F(pop)$ 为该个体的适应度函数, 它由降维率 $R(pop)$ 和复现精度 $M(pop)$ 两部分构成。式(7)中, sigmoid函数^[24]是神经网络常用的激活函数, 其表达式如式(9)

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-4x}} \quad (9)$$

式中, α 是缩放系数, 用以改变sigmoid函数的变化率。 m 是原始特征集的特征总数, m' 是种群个体包含的特征数, 易得 $\frac{(m - 2m')}{m} \in [-1, 1]$ 。

以 $\frac{(m - 2m')}{m}$ 的值为自变量, 得到式(7)在不同 α 下的函数图像如图4所示。

图 4 式(7)在不同 α 下的函数图像

由图 4 可知, 当 $\alpha < 1$ 并不断减小时, sigmoid 函数在 $[-1, 1]$ 区间内的变化率放缓, 通过增加降维率获得的适应度函数收益减少, 因此遗传过程更倾向于选择复现精度高的个体; 当 $\alpha > 1$ 时, 在一定范围内(如 $\alpha = 2$, $[-0.4, 0.4]$), sigmoid 函数在该区间内变化率增大, 通过增加降维率获得的适应度函数收益增大, 遗传过程更容易选择具有较高降维率的个体, 超过该范围则更倾向于选择复现精度高的个体, 从而很大程度上避免了算法收敛到降维率 100% 的情况。

式(8)是 RBFNN 复现精度的计算公式, 其中 $\hat{y}_i$ 是第 i 个样本的网络输出值, y_i 是第 i 个样本的真实值, n 为样本个数。 s 和 d 是由 RBFNN 复现精度的上下限得到的参数, 分别由式(10)和(11)计算得到

$$s = \frac{|ub - lb|}{2} \quad (10)$$

$$d = \frac{ub + lb}{2} \quad (11)$$

式中, ub 为 RBFNN 复现精度的上限, lb 为复现精度的下限。对于中低维特征集($m \leq 100$)可取 $ub=8 \sim 10$, $lb=4$, 其中 ub 的值对应的是 50 维数据集下, 二次函数和线性函数的平均复现精度, lb 的值是 50 维条件下带噪声特征或复杂非线性函数(如符号函数)的平均复现精度。对于高维特征集($m > 100$)可取 $ub=4 \sim 6$, $lb=2$, 由 250 维的对应平均复现精度得到。 ub 、 lb 也可以通过式(12)、(13)确定

$$ub = \max_{j=1}^m \left(-\lg \left(\frac{1}{n} \sum_{i=1}^n (\hat{y}_{ij}(\mathbf{x}_{i, \bar{j}}) - y_{ij})^2 \right) \right) \quad (12)$$

$$lb = \min_{j=1}^m \left(-\lg \left(\frac{1}{n} \sum_{i=1}^n (\hat{y}_{ij}(\mathbf{x}_{i, \bar{j}}) - y_{ij})^2 \right) \right) \quad (13)$$

其中, $\mathbf{x}_{i, \bar{j}} = [x_{i1}, x_{i2}, \dots, x_{i, j-1}, x_{i, j+1}, \dots, x_{im}]$, $i=1, 2, \dots, n$, 其元素由第 i 个样本去掉第 j 个特征的其他全部特征

组成, $\hat{y}_{ij}$ 和 y_{ij} 分别是由输入 $\mathbf{x}_{i, \bar{j}}$ 得到的 RBFNN 的估计值和真实值。

遗传算法的交叉过程采用随机产生交叉位置的方法对父代与母代染色体进行交换, 变异过程依照伯努利分布随机产生变异位置, 并把该位置的二进制编码取反, 为了避免陷入局部最小, 在每次迭代中逐渐增大变异率, 直到迭代结束。

GA-RBFNN 算法伪代码如下所述, 其中输入参数的对应含义如下: NPop 为种群大小, Niter 为世代数, RCross 为杂交率, RSelect 为选择率, RVar 为变异率。

输入: 制造样本 $\mathbf{X} \in R^{n \times m}$, alpha, NPop, Niter, RCross, RSelect, RVar

输出: 制造特征集合 $\mathcal{S}$

- 1 归一化制造样本集 $\mathbf{X}$;
- 2 划分训练样本集 $\mathbf{X}_{\text{train}}$ 、验证集 $\mathbf{X}_{\text{val}}$ 、测试集 $\mathbf{X}_{\text{test}}$;
- 3 以式(10)~(13)计算 ub, lb, d, s ;
- 4 初始化种群矩阵 $\mathbf{Pop} \in \{0, 1\}^{NPop \times m}$;
- 5 while ($i \leq \text{Niter}$) do:
- 6 for $i = 1$ to NPop do
- 7 以 $\text{Pop}(i)$ 编码代表的特征作为 RBFNN 的输入, 以全部特征作为输出, 利用 $\mathbf{X}_{\text{test}}$ 训练 RBFNN;
- 8 以 $\mathbf{X}_{\text{test}}$ 计算复现精度;
- 9 以式(6)~(8)计算个体适应度函数;
- 10 end for
- 11 以 RSelect 的选择率及轮盘赌选择父代个体, 且必定保留该次迭代的最优个体;
- 12 以 RCross 概率发生交叉;
- 13 以 RVar 概率发生变异;
- 14 $i=i+1$;
- 15 $\text{RVar}=0.1+0.1*i/\text{Niter}$;
- 16 end while
- 17 取 $\mathcal{S}$ 为最优个体编码所代表的制造特征子集;

2.2.3 高维问题中的算法变体

由于 RBFNN 是基于高斯核函数的神经网络, 而高斯核函数与样本间距离有关, 因此对于样本维度过大的情况, RBF 神经网络通常不能获得较高的拟合精度。

图 5 是 RBFNN 的复现精度与输入端维度的关系图。设输入端特征 $f_1, f_2, \dots, f_k$, 其中 $f_i, f_j, i \neq j$ 为相互独立的正态分布变量, 通过不断引入新的变量增加输入端的维度。输出端统一为特征 f_1 。由图 5 可知, 对于同一变量的复现精度随着不相关输入维度的增多, 呈指数增长。

图5 输入端维度对 RBFNN 复现精度的影响

针对这一问题,提出解决方案如下。考虑输出端增维对于复现某一固定特征的影响,设计以下实验。设输入端特征变量 $f_1, f_2, \dots, f_k$, 其中 k 为定值。输出端特征变量取 $f_1, f_2, \dots, f_i, i=1, 2, \dots, k$ 分别进行 k 次网络复现,并计算第 1 个维度的复现精度,其结果如图 6 所示。由图 6 可知,不相关输出变量对于单一固定输出变量的复现精度没有影响。

图6 输出端维度对 RBFNN 复现精度的影响

利用这一特性,可以通过下式预筛选出可能存在冗余性的特征集

$$\mathcal{R} = \left\{ f_j \mid -\lg \left(\frac{1}{n} \sum_{i=1}^n (\hat{y}_{ij}(x_i, \hat{f}_j) - y_{ij})^2 \right) \geq T, j=1, 2, \dots, m \right\}$$

式中, T 为可以接受的复现精度阈值,可取 2~4。将 $\mathcal{R}$ 中所包含的特征作为网络输出端的特征。为了避免神经网络学习到全等表示,输入端的候选集合为 $\{\mathcal{F} - \mathcal{R}\} \cap \mathcal{R}^*$, 其中 $\mathcal{F}$ 为制造特征全集, $\mathcal{R}$ 为冗余特征集, $\mathcal{R}^*$ 为施加稀疏性约束的冗余特征集,如稀疏性为 0.05 时, $\mathcal{R}$ 中的特征有 5% 的几率被选入输入端的候选特征集合中。与此同时,降维率不必再作为适应度函数的一个优化目标,将式(8)从式(6)中删除。

根据这一约束,对 GA-RBFNN 算法(第 2.2.2 节)

的程序步骤 4 进行修改,使得初始化得到的个体在与 $\mathcal{R}$ 对应的维度上满足稀疏性约束。同时在程序步骤 13 与程序步骤 14 之间加入子代个体的稀疏性检查,保证个体满足稀疏性约束。该方法减小了输入端的维度,避免 RBFNN 在高维特征选择问题中的精度下降。

经过一阶段 GA-RBFNN 算法后,数据集内非线性冗余特征被基本删除。在剩余无冗余特征中进行基于 LS 的特征评价方法,可以很大程度上避免重复选择具有相近相关性的特征。

2.3 基于 LS 的第二阶段特征选择方法

拉普拉斯得分(LS)是一种基于拉普拉斯特征映射^[25]及局部保留投影^[26]的过滤式特征选择方法。它基于两个基本假设:一是空间中具有相近距离的两个点很可能属于同一类别,二是空间点的局部结构比全局结构更有效。因此在完全无标签的样本中,LS 能够通过评价特征的局部保留能力选择特征,从而筛选出最能分辨数据局部结构的特征。LS 的一般步骤如下所述。

(1) 构建样本的最邻近图 G。

当两个样本 $x_i, x_j, i \neq j$ 较“近”时,将两个样本通过边相连。可以取 x_i 的 k 邻近点,建立最近邻近图。

(2) 计算相似度矩阵 S。

若两个点通过边相连,计算该边的权重为

$$s_{ij} = \exp \left(-\frac{\|x_i - x_j\|^2}{t} \right), \text{ 否则权重为 } s_{ij} = 0. \text{ 则}$$

$$S = [s_{ij}]_{m \times m}.$$

(3) 计算图的拉普拉斯矩阵 L。

设特征 $f_i = [f_{i1}, f_{i2}, \dots, f_{in}]^T$, 其中 f_{ij} 为第 i 个特征在第 j 个样本上的取值。计算度矩阵 $D = \text{diag}(S_1)$, 其中 $\text{diag}(X)$ 为由矩阵 X 对角元素构成的对角矩阵, $\mathbf{1} = [1, 1, \dots, 1]^T$, 则 $L = D - S$ 。

(4) 计算第 i 个特征的拉普拉斯得分 L_i

$$L_i = \frac{\tilde{f}_i^T L \tilde{f}_i}{\tilde{f}_i^T D \tilde{f}_i} \quad (14)$$

$$\tilde{f}_i = f_i - \frac{f_i^T D f_i}{\mathbf{1}^T D \mathbf{1}} \mathbf{1}$$

拉普拉斯得分计算公式是由特征样本间相似度之和与该特征样本间的方差之比得到的,因此越小的 L_i 意味着该特征维度下类内样本之间距离越小,同时类间样本之间距离越大。在一阶段消除特征冗余性的基础上,运用拉普拉斯的分能够更好地拉开特征之间的得分差异,从而更好地筛选出与分类有关的关键特征。

3 试验与分析

试验部分首先通过 UCI Iris 数据集^[27]设计一阶段算法有效性试验。在 Iris 原有四个互不相关的特征基础上进行非线性拓展, 增大数据集的非线性冗余性。通过对拓展数据集应用 GA-RBFNN 算法, 并与基于正则自表示^[9] (Regulized self representation, RSR) 的去冗余特征选择算法对比, 验证一阶段算法对于去除数据集非线性冗余性的有效性。随后在 UCI Secom 的高维故障诊断数据集^[29]上应用完整的两阶段特征选择算法, 采用分类准确性评价特征选择算法的性能, 并与 LS、RSR 等无监督特征选择算法作对比, 验证了本文提出的方法在数据去冗余能力及改善分类效果两方面均有其优越性。

试验 PC 的基本参数为: Inter Core i7-7700K, 主频 4.20 GHz, 内存 16 GB, 试验软件为 Matlab 2017a。

3.1 一阶段算法有效性试验

第一个试验采用公开数据集人工合成后的数据集对算法去除非线性冗余性的能力进行验证。试验采用的数据集是 UCI Iris 数据集, 其包含四个不相关的特征, 总计 150 个样本。该数据集特征维度小, 样本少, 因此, 算法验证时间短, 并且特征间相关性差异大, 常被特征选择相关文献^[9]用作验证算法有效性的试验数据集。利用常规非线性函数和高斯噪声生成原四维特征的非线性组合特征, 得到 15 维拓展 Iris 数据集, 各维度生成方式如表 2 所示。

表 2 拓展 Iris 数据集各维度生成方式

序号	函数性质	表达式
1~4	数据集原始特征	f_1, f_2, f_3, f_4
5	线性冗余特征	$f_1 + f_2$
6	线性冗余特征	$f_2 - f_3$
7	含噪声冗余特征	$f_3 + f_4 + 0.05\varepsilon$
8	含噪声冗余特征	$f_1 + f_2 + f_3 + f_4 + 0.05\varepsilon$
9	反三角函数	$\text{Arctan}(f_1)$
10	对数函数	$\ln(f_2)$
11	幂函数	$(f_1)^2$
12	指数	e^{f_4}
13	三角函数	$\sin(f_1 + f_2)$
14	复杂非线性函数	$\text{sgn}(f_3 + f_4 - f_2)$
15	复杂非线性函数	$\text{sigmoid}(f_1 + f_2 + f_3 + f_4)$

试验采用去冗余性常用的 RSR 算法作为对比算法, 其对应的优化问题可表示为

$$W = \arg \min_w (\|X - XW\|_{2,1} + \lambda \|W\|_{2,1}) \quad (15)$$

由式(15)可知, 通过 RSR 算法过滤掉的特征仅是源特征集中的线性冗余特征。试验中取 RSR 的正则化系数 $\lambda=0.1$ 。

两阶段特征选择算法的试验参数取 $\alpha=1$, $ub=10$, $lb=4$, $NPop=60$, $Niter=100$, $RCross=0.7$, $RSelect=0.6$, $RVar=0.1$ 。

试验结果如图所示。图 7 为 RSR 权值矩阵 W 的热力图, 颜色越深代表该次表达过程中的特征得分越高。对 W 行进行求和得到 15 个特征(热力图的行)在 15 次表达(热力图的列)中的累积得分, 如图 8 所示。由图中得出, RSR 对于线性冗余性的去除能力较强, 但完全无法去除非线性冗余特征。

图 7 RSR 权值矩阵元素数值热力图

图 8 RSR 特征选择结果图

GA-RBFNN 算法的最终收敛结果是 4 维剩余特征。为可视化试验结果, 且便于与 RSR 比较, 通过对遗传算法完成所有时代后种群中的个体各维度编码求和取平均, 得到与 RSR 类似的特征得分评价图如图 9 所示。其中排除基因突变产生的特征 15, 算法选择出的 4 维基础特征能够完全代表 15 维特征, 由此证明本文设计的一阶段算法能够有效去除数据集的非线性冗余特征及识别具有噪声的冗余特征。

图9 GA-RBFNN 特征选择结果图

3.2 两阶段算法有效性试验

第二个试验采用实际数据集对完整的两个阶段算法提升分类效果的能力进行验证。试验采用的数据集为 UCI Secom 产品质量检测数据集, 包含 1 567 个样本和 591 个质量相关特征。

从特征的角度分析, 该数据集属于高维数据集, 它的每一维是由半导体制造过程中的各传感器采集而来或在制造过程中测量得到的产品质量相关特征, 并非所有的特征对最终的产品质量检测均有帮助。这些特征中除了有用特征外, 还包含无用特征、不相关特征、冗余特征及噪声特征。从样本的角度分析, 该数据集包含无效值(如 NaN)、离群值噪声等。此外 1 567 个样本分为合格样本与不合格样本, 其中合格样本有 1 463 个, 不合格样本有 104 个, 属于不平衡数据集。

按照文献[3]中的处理方法对该数据集进行简单预处理, 包括缺失值处理、无用特征删除、平衡化以及归一化。经过以上四个步骤, 剩余样本 209 个, 其值均在[0,1]范围内。剩余特征 287 维。由于 RBFNN 需要较多的样本进行训练, 并且一阶段只消除冗余性, 数据集是否平衡对于该目标影响不大, 因此 RBFNN 训练样本依旧采用经过缺失值处理和无用特征删除后的包含 1 480 个样本与 287 个特征的数据集。

3.2.1 评价指标

由于方法的目的是去除数据集冗余性及提升分类精度, 故按照文献的一般定义方法分别对数据集的剩余冗余度(RED)^[10]及分类准确率(ACC)^[9]进行定义

$$\text{RED}(f^s) = \frac{1}{m(m-1)} \sum_{f_i, f_j \in f^s, i > j} \rho_{ij} \quad (18)$$

式中, f^s 为剩余特征集合; ρ_{ij} 为第 i 个特征与第 j 个特征间的相关系数。

$$\text{ACC} = \frac{\sum_{i=1}^n \delta(p_i, m(q_i))}{n} \quad (19)$$

其中 $\delta(x, y)$ 满足 $x = y$ 时, $\delta(x, y) = 1$; $x \neq y$ 时, $\delta(x, y) = 0$ 。 $m(q_i)$ 是采用最优分类算法在训练样本 q_i 上的分类类别输出值。对于一阶段算法, 期望得到较小的 RED 值; 对于二阶段算法, 期望得到较大的 ACC 值。

此外, 定义 RBFNN 的还原精度即均方差(Mean square error, MSE)作为辅助评价指标度量筛选后的特征保留原数据集信息能力的大小, 其表达式如下

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (20)$$

式中, $\hat{y}_i$ 为 RBFNN 对第 i 个样本的预测值; y_i 为样本真实值。

3.2.2 试验参数设定及试验结果

考虑到神经网络需要的样本较多, 采用第 3.2.1 节获得的 1 480 个样本数据集进行一阶段的试验。试验采用高维变体 GA-RBFNN 方法, 稀疏度分别设定为 0.1、0.2、0.4, $ub=6$, $lb=3$, $NPop=600$, $Niter=100$, $RCross=0.7$, $RSelect=0.6$, $RVar=0.1$ 。RSR 取正则化系数 λ 取值于集合 {0.01, 0.1, 1, 10, 100}, 并根据 GA-RBFNN 的收敛结果, 依据 RSR 特征得分选择与 G 相同数目的特征组成 RSR 的特征集。

GA-RBFNN 与 RSR 的特征选择结果分别如表 3 与表 4 所示。两种特征选择算法剩余冗余度与原始数据冗余度的对比结果见表 5, 其中 RSR 取最优结果。

表3 GA-RBFNN 特征选择结果表

稀疏性	特征数	MSE	RED
0.1	195	0.096 5	0.011 31
0.2	214	0.049 5	0.013 91
0.4	231	0.007 885	0.015 53

表4 RSR 特征选择结果表

正则化系数	特征数	MSE	RED
0.01	195	1.981	0.013 649
	214	0.927	0.016 149
	231	0.402	0.018 835
0.1	195	4.812	0.013 938
	214	0.836	0.016 382
	231	0.220	0.018 901
1	195	2.188	0.013 787
	214	0.267	0.016 246
	231	0.139	0.018 606
10	195	0.523	0.013 638
	214	0.467	0.016 213
	231	0.314	0.019 377
100	195	4.300	0.014 518
	214	1.011	0.017 059
	231	0.376	0.019 303

表 5 GA-RBFNN 与 RSR 对比结果表

特征数	RSR	GA-RBFNN	原始冗余度
195	0.013 638	0.011 31	
214	0.016 149	0.013 91	0.026 927
231	0.018 606	0.015 53	

由表 5 中结果, GA-RBFNN 能够有效降低原始数据集中的冗余性, 其效果优于 RSR 算法。

在一阶段已选特征的基础上, 以 195 个特征为例, 进一步进行第二阶段特征选择。通过 LS 算法分别对 195 个剩余特征的拉普拉斯得分进行计算, 并按升序排列。分别选择拉普拉斯得分最低的前 50 个、前 60 个直到前 150 个, 作为分类算法的输入特征, 进行 11 组试验。每组试验中, 考虑到样本数量较少, 采用 10 折交叉验证的方法验证分类效果, 并重复进行 10 次试验, 消除偶然性。

分类算法统一采用 Matlab 中分类工具箱中的 SVM 算法, 每次试验调整 SVM 的 kernel scale 直至获得最优的分类效果作为分类准确性 ACC 的试验结果。

为了验证消除了非线性冗余性后的输入特征比不消除冗余特征、直接评估拉普拉斯得分得到的输入特征分类效果更好, 设计与 LS 算法进行对比试验。为了验证两阶段算法 TSFS 的效果比其他无监督特征选择算法分类效果更好, 设计与 RSR 算法进行对照试验。试验结果如表 6 所示。

表 6 各算法分类精度对比表

特征数	LS(%)	RSR(%)	TSFS(%)
50	58.46	60.48	62.43
60	59.42	63.12	63.34
70	62	65.22	65.06
80	62.97	65.22	66.93
90	62.83	65.32	67.09
100	64.27	66	67.35
110	66.99	67.36	67.67
120	67.62	66.98	67.87
130	67.18	66.2	68.43
140	66.69	65.14	69.25
150	67.72	65.44	70.15
无特征选择		67.5	

为进一步可视化试验效果, 对表 6 结果绘制散点图, 并利用 3 次多项式逼近散点, 如图 10 所示。从图中可以看出, 相同维度下, TSFS 的分类准确率均优于其他两种算法, 当维度大于 110 维时, TSFS 的分类准确率超过了不进行特征选择的分类准确率, 说明该方法对于提高分类准确性有一定的帮助。LS 较 TSFS 的准确度波动性更大, 可能是由于引入

了冗余变量导致无关特征维度增多使得分类精度下降。RSR 在这个数据集上对于改善分类效果的作用不是很明显。

图 10 各算法分类精度对比图

综上所述, 本文设计的两阶段特征选择算法较 RSR 算法能更有效地降低原始数据集的冗余度, 与 LS 算法相比能更有效地提高数据的价值密度, 并有效改善分类精度。

4 结论

针对制造过程数据无标签、高冗余性和噪声沿特征分布的特点, 以降低数据集冗余性、提升数据价值密度、改善分类效果为目的, 提出了两阶段无监督特征选择算法。基于 RBFNN 可以拟合非线性关系的特性, 提出利用低维特征集作为 RBFNN 的输入表示全部特征的一阶段去冗余方法, 并与 LS 相结合, 在无冗余子集中选择出能够保持数据局部特性的特征, 从而达到改善分类效果的目的。通过在合成数据集与实际数据集上验证, 并与 LS 与 RSR 算法作对比, 验证了本文提出的方法去除冗余性的能力更强、分类效果更优。

今后的工作将以降低该方法复杂度, 缩短计算时间为重点进行。此外, 还将着重研究算法参数的自适应选择方法。

参 考 文 献

- [1] 张洁, 秦威, 鲍劲松. 制造业大数据[M]. 上海: 上海科学技术出版社, 2016.
ZHANG Jie, QIN Wei, BAO Jingsong. Big data in manufacturing industry[M]. Shanghai: Shanghai Scientific and Technical Publishers, 2016.
- [2] 杨文献, 姜节胜. 基于数据挖掘的柴油机气门故障诊断技术研究[J]. 机械工程学报, 2004, 40(10): 25-29.

- YANG Wenxian, JIANG Jiesheng. Data mining based technique for diagnosing the faults presenting on engine valves[J]. Chinese Journal of Mechanical Engineering, 2004, 40(10): 25-29.
- [3] 闫伟. 基于数据挖掘的复杂产品关键质量特性识别的方法研究[D]. 天津: 天津大学, 2012.
- YAN Wei. On Critical-to-quality characteristics identification for complex products using data mining[D]. Tianjin: Tianjin University, 2012.
- [4] KEOGH E, MUEEN A. Curse of dimensionality[J]. IND. ENG. Chem., 2017, 29(1): 48-53.
- [5] 徐峻岭, 周毓明, 陈林, 等. 基于互信息的无监督特征选择[J]. 计算机研究与发展, 2012, 49(2): 372-382.
- XU Junling, ZHOU Yuming, CHEN Lin, et al. An unsupervised feature selection approach based on mutual information[J]. Journal of Computer Research and Development, 2012, 49(2): 372-382.
- [6] WANG Junliang, ZHANG Jie. Big data analytics for forecasting cycle time in semiconductor wafer fabrication system[J]. International Journal of Production Research, 2016, 54(23): 7231-7244.
- [7] WANG Junliang, ZHANG Jie, WANG Xiaoxi. A data driven cycle time prediction with feature selection in a semiconductor wafer fabrication system[J]. IEEE Transactions on Semiconductor Manufacturing, 2018, 31(1): 173-182, 5.
- [8] ISLAM M R, UDDIN M S, KHAN S, et al. Multi-core accelerated discriminant feature selection for real-time bearing fault diagnosis[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer International Publishing, 2016: 645-656.
- [9] ZHU P, ZUO W, ZHANG L, et al. Unsupervised feature selection by regularized self-representation[J]. Pattern Recognition, 2015, 48(2): 438-446.
- [10] ZHAO Z, WANG L, LIU H. Efficient spectral feature selection with minimum redundancy [C]// Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July. DBLP, 2010.
- [11] JOHN G H, KOHAVI R, PFLEGER K. Irrelevant features and the subset selection problem[J]. Precision Engineering, 2004, 2(4): 314-318.
- [12] ZHAO Z, LIU H. Spectral feature selection for supervised and unsupervised learning[C]// Machine Learning, Proceedings of the Twenty-Fourth International Conference. DBLP, 2007: 1151-1157.
- [13] ZHAO Z, WANG L, LIU H. Efficient spectral feature selection with minimum redundancy [C]// Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July. DBLP, 2010.
- [14] CAI D, ZHANG C, HE X. Unsupervised feature selection for multi-cluster data [C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 333-342.
- [15] 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 Xu, WANG Xiaodan, ZHANG Yuxi, et al. Summary of feature selection algorithms[J]. Control and Decision, 2012, 27(2): 161-166.
- [16] ZHANG X L, CHEN W, WANG B J, et al. Intelligent fault diagnosis of rotating machinery using support vector machine with ant colony algorithm for synchronous feature selection and parameter optimization[J]. Neurocomputing, 2015, 167: 260-279.
- [17] CERRADA M, SÁNCHEZ R V, CABRERA D, et al. Multi-stage feature selection by using genetic algorithms for fault diagnosis in gearboxes based on vibration signal[J]. Sensors, 2015, 15(9): 23903-23926.
- [18] TABAKHI S, MORADI P, AKHLAGHIAN F. An unsupervised feature selection algorithm based on ant colony optimization[J]. Engineering Applications of Artificial Intelligence, 2014, 32: 112-123.
- [19] GUYON I, ELISSEEFF A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 3: 1157-1182.
- [20] HOQUE N, BHATTACHARYYA D K, KALITA J K. MIFS-ND: A mutual information-based feature selection method[J]. Expert Systems with Applications, 2014, 41(14): 6371-6385.
- [21] PARK J, SANDBERG I W. Universal approximation using radial-basis-function networks[J]. Neural Computation, 1991, 3(2): 246-257.
- [22] WHITLEY D. A genetic algorithm tutorial[J]. Statistics and Computing, 1994, 4(2): 65-85.
- [23] LEONTARITIS I J, BILLINGS S A. Model selection and validation methods for non-linear systems[J]. International Journal of Control, 1987, 45(1): 311-341.
- [24] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436.
- [25] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C]// International Conference on Neural Information Processing Systems: Natural and Synthetic. MIT Press, 2002: 585-591.
- [26] HE X, NIYOGI P. Locality preserving projections[J]. Advances in Neural Information Processing Systems, 2003, 16(1): 186-197.
- [27] HE Xiaofei, DENG Cai, PARTHA N. Laplacian score for

feature selection[J]. Advances in neural Information Processing Systems, 2006: 507-514.

作者简介: 张洁(通信作者), 女, 1963 年出生, 教授。主要研究方向为制造系统的建模、优化调度与控制、制造信息工程和智能制造系统。

E-mail: mezhangjie@dhu.edu.cn

盛夏, 男, 1993 年出生。主要研究方向为数字化制造与智能制造。

E-mail: shengxia0111@sjtu.edu.cn

张朋, 男, 1988 年出生, 博士研究生。主要研究方向为制造系统优化调度与控制

E-mail: pgnahz@qq.com

秦威, 男, 1985 年出生, 副教授。主要研究方向为复杂(制造)系统的建模、控制与优化, 制造大数据分析, 制造智能、人工智能在制造业的应用。

E-mail: wqin@sjtu.edu.cn

赵新明, 女, 1960 年出生, 副教授。主要研究方向为机械设计及图形处理技术。

E-mail: zhaoxm@sjtu.edu.cn