文章编号:1001-4098(2019)01-0134-09

基于多目标鲸鱼优化的关键质量特性识别方法*

李岸达[1],何 桢[2],王 庆[1]

(1.天津商业大学管理学院,天津 300134;2.天津大学管理与经济学部,天津 300072)

摘 要:提出基于多目标鲸鱼优化(multi-objective whale optimization algorithm, MWOA)的产品关键质量特性(特征)识别方法。首先,针对非平衡制造过程数据,将识别模型构建为最大化 G-mean 指标和最小化特征数的多目标特征选择问题。其次,提出群智能优化算法 MWOA 求解模型。MWOA 针对特征选择提出一种新的多样性帕累托排序策略,能够对解进行优劣排序的同时保证群体多样性。同时,在 MWOA 中嵌入变异操作以解决鲸鱼优化易陷入局部最优的缺点。选取 3 组非平衡制造过程数据验证算法有效性。实验结果表明,所提算法在数据非平衡条件下能够有效识别关键质量特性。

关键词:关键质量特性识别;特征选择;多目标优化;鲸鱼优化算法;非平衡数据

中图分类号:F273.2 **文献标识码**:A

1 引言

产品加工制造过程中包含若干过程参数、产品尺寸参数等质量特性,这些质量特性是影响产品整体质量水平的重要因素[1,2]。然而,不是所有质量特性都是导致产品质量缺陷的关键因素。因此,识别影响产品质量的关键质量特性(key quality characteristics, KQC),进而对所识别 KQC 进行参数优化,是产品质量持续改进的基础[3]。随着科技的发展,产品结构呈现不断复杂化的趋势,产品质量特性数急剧增长,显著提升了从生产线所收集制造过程数据的维度,导致数据驱动的传统多元线性回归等方法难以用于这类高维数据进行 KQC 识别[4]。机器学习领域的特征选择算法能有效针对高维数据进行特征降维,近年来基于特征选择的 KQC 识别研究受到广泛关注[3,5-9]。

特征选择的主要目标是选择与类标签(产品质量)相关的关键特征并过滤无关、冗余特征,通常被定义为最大化特征(质量特性)对类标签的预测能力和最小化所选择特征数(质量特性数)的多目标优化问题[10]。由于特征选择是典型的 NP-hard 问题,启发式算法通常被用于求解该问题[11,12]。最典型的两个方法是基于爬山策略的序列向前(sequential forward selection, SFS)和序列向后(sequential backward selection, SBS)算法[11]。SFS 从一个空的特征子集开始,逐步选择能使特征子集有更好预测能力的特征,直到特征子集预测能力不能进一步提高。与 SFS 相反,SBS 从完整的特征子集开始,逐步剔除无效特征使特征子集预测能力提高。SFS 和 SBS 基于爬山策略,其缺点是易陷入局部最优。由于进化算法、群智能优化算法(如遗传算法、粒子群优化)有不错的广域搜索能力,适用于求解 NP-hard 问题,基于该类算法的特征选择是近年来的研究热点。文献[13-16]提出基于单目标遗传算法、粒子群优化的特征选择算法。为了求解多目标特征选择模型,该类方法建立综合的适应度函数将多目标问题转换为单目标问题。但是,建立综合适应度函数时各分目标权重不易确定是其缺点。近年来,多目标进化和群智能优化算法如 NSGA-II、CMDPSO、NSPSO[8,10,17] 被用于建立特征选择算法,该类方法有效避免了单目标优化算法不易确定分目标权重的问题。然而,制造过程数据通常是合格产品多于不合格产品的非平衡数据,这些方法在建立特征选择算法时没有考虑数据非平衡性,不适用于直接用于

* **收稿日期**:2018-11-29;**修订日期**:2018-12-27

基金项目:教育部人文社会科学研究一般项目(19YJC630071);国家自然科学基金(71661147003,71532008);天津市教委科研计划项目成果(161072,基于非平衡数据的复杂产品关键质量特性识别研究)

作者简介:李岸达(1989-),男,博士,讲师,研究方向为质量工程、智能算法;何桢 (1967-),男,博士,教授,研究方向为质量工程与质量管理、工业工程。

KQC 识别。

Mirjalili 和 Lewis[18] 于 2016 年提出一种新的群智能优化算法——鲸鱼优化算法（whale optimization algorithm，WOA）。该方法受座头鲸的泡泡网觅食（bubble－net attacking）行为启发，建立了充分结合探索能力和开发能力的位置（解）更新策略。WOA 具有原理简单、人工设置参数少、收敛速度优等优点，已在大规模优化[19]、神经网络参数优化[20]、热电经济调度[21]等领域得到有效应用。由于特征选择问题具有高维度、耗时多的特点，优化算法能否快速收敛直接影响特征选择的时间有效性，这些特点使得 WOA 算法很适用于特征选择问题。因此，Mafarja 等[22, 23]、Sharawi 等[24] 将 WOA 用于特征选择，取得了不错效果。然而，WOA 是单目标优化算法，同样存在不易确定分目标权重的缺点。据作者所知，目前没有研究提出多目标 WOA 用于特征选择，基于多目标 WOA 的特征选择算法值得进一步研究。

为了识别产品 KQC，本文提出基于多目标鲸鱼优化（multi－objective whale optimization algorithm，MWOA）的特征选择算法。由于制造过程数据是合格产品（样本）较多、不合格产品（样本）较少的非平衡数据，使用传统特征选择模型会导致有偏的特征选择结果，因此首先构建针对非平衡制造过程数据的多目标特征选择模型。为了求解该模型，提出 WOA 的多目标优化版本 MWOA。MWOA 采用一种新的多样性帕累托排序策略，并嵌入变异操作，以保证算法的优化效率。实验结果表明，所提方法能够基于非平衡制造过程数据有效进行 KQC 识别。

2　KQC 识别问题

设 $F=\{f_1,f_2,\cdots,f_N\}$ 为产品生产制造过程中的 N 个质量特性（特征），$Q\in\{-1,1\}$ 为产品质量水平（类标签），其中负类“－1”表示产品合格，正类“1”表示产品不合格。制造过程数据集可表示为 $D=[E,C]$，其中 $E\in\mathbb{R}^{M\times N}$ 和 $C\in\mathbb{R}^{M}$ 分别表示 M 个样本的特征观测值和类标签观测值。KQC 识别可以定义为选择一个特征子集 $F_s\subset F$ 使得特征重要性度量 $J(F_s)$ 最大化和特征数 $\#F_s$ 最小化的多目标特征选择问题。通常情况下，生产过程中的合格产品显著多于不合格产品，导致 D 为非平衡数据。因此，本文假定所处理数据集为非平衡数据，并在此基础上建立特征重要性度量指标 $J(F_s)$。

分类精度（accuracy）是特征选择中常用的特征重要性度量指标。表 1 为二分类问题的混淆矩阵，#TP、#FN、#FP、#TN 分别表示分类器得到的真正、假负、假正、真负样本的数量，则分类精度定义如下：

$$\text{Accuracy}=\frac{\#\text{TP}+\#\text{TN}}{\#\text{TP}+\#\text{TN}+\#\text{FP}+\#\text{FN}} \tag{1}$$

然而，对于非平衡的制造过程数据集，多（负）类样本对分类精度的影响远高于少（正）类样本，高的分类精度并不能有效反映对少类样本的分类效果[25]。在数据非平衡条件下，仍然采用分类精度度量特征子集重要性会导致有偏的特征选择结果。

针对非平衡数据，本文采用敏感性（Sensitivity）和特异性（Specificity）的几何平均 G－mean 指标代替分类精度作为特征重要性度量，即

$$J(F_s)=\text{G}-\text{mean}=\sqrt{\text{Sensitivity}\times\text{Specificity}} \tag{2}$$

其中，敏感性和特异性分别代表对正类样本和负类样本的分类精度，定义如下：

$$\text{Sensitivity}=\frac{\#\text{TP}}{\#\text{TP}+\#\text{FN}} \tag{3}$$

$$\text{Specificity}=\frac{\#\text{TN}}{\#\text{TN}+\#\text{FP}} \tag{4}$$

由式(2)可知，大的 G－mean 值要求敏感性和特异性值都较大，优化该指标能够保证对正负类样本都有高的分类精度，从而不易受到数据非平衡所带来的影响。综上所述，本文定义 KQC 识别为最大化 $J(F_s)=\text{G}-\text{mean}$ 和最小化 $\#F_s$ 的多目标优化特征选择问题。为了便于求解，将最大化 $J(F_s)$ 转化为最小化 $1-J(F_s)$，因此得到如下多目标特征选择模型：

$$\begin{aligned}&\min f_1(F_s)=1-J(F_s)\\&\min f_2(F_s)=\#F_s\\&\text{s.t. } F_s\subset F\end{aligned} \tag{5}$$

对于公式(5)所定义模型，目标函数 $f_2(F_s)$ 通过直接计算 F_s 中的特征数得到，$f_1(F_s)$ 实际上是一种分类指标，需要使用分类器来进行估算。本文采用特征选择所经常采用的 Wrapper 框架，使用内部 5 折交叉验证法于训练数据集估计得到 $f_1(F_s)$，具体步骤详见文献[11]。

表 1　混淆矩阵

		预测类别	
		正(1)	负(－1)
真实类别	正(1)	#TP	#FN
	负(－1)	#FP	#TN

3　多目标鲸鱼优化算法(MWOA)

3.1　算法步骤

本节提出 WOA 的多目标版本 MWOA 求解式(5)所

定义多目标特征选择模型。所提 MWOA 算法步骤如算法 1 所示。MWOA 中采用了多样性帕累托排序对解进行排序，该方法在采用帕累托法则对解排序的同时，针对特征选择问题嵌入了群体多样性保持策略，具体见 3.3 节。算法第 4 到 8 步为解的更新策略，通过模拟座头鲸泡泡网觅食行为中的包围捕食策略(encircling prey)、螺旋游走策略(spiral updating position)或猎物搜寻策略(search for prey)对群体中的解进行迭代进化。但是，标准 WOA 算法更新策略的缺点是易早熟而使算法陷入局部最优[19]。为了解决该问题，在算法第 9 步引入一种变异操作，以使算法逃离局部最优点。另外，标准 WOA 算法在优化过程中，无条件接受更新策略所产生的新解。但是这种策略不能保证在迭代过程群体中解的质量逐步稳定提高，影响了算法优化效率。为了解决此问题，在算法第 11 步对更新策略产生的新解与上代群体中的解进行整体性排序，并选择前 N^p (群体规模)个解作为新一代群体。最后，在算法第 13 步引入多目标决策常用策略——理想点法，从所得非支配解中选择能够平衡目标函数 f_1 和 f_2 的最佳调和解(详见文献[8])。以下为算法各部分详细内容。

算法 1 MWOA 算法

1.初始化鲸群位置(解) $X_i, i=1,\cdots,N^p$，N^p 为群体规模，令 $t=0$，最大迭代次数为 $t_{\max}$；

2.得到各解的适应值 $\{f_1(X_i), f_2(X_i)\}$，将群体中的解和适应值 $[X_i, f_1(X_i), f_2(X_i)], i=1,\cdots,N^p$ 添加到集合 Ω_t；

3.利用多样性帕累托排序对 Ω_t 中的解进行排序，将非支配解添加到 *Leader_set*；

While($t < t_{\max}$)

4.令 $\Gamma = \Omega_t$

For each X_i in Ω_t

5.更新参数 A_i, C_i, l_i, p_i (见 2.2、2.3 节)

If1 $p_i < 0.5$

If2 $|A_i| < 1$

6.采用包围捕食策略更新 X_i；

Else If2 $|A_i| \geq 1$

7.采用猎物搜寻策略更新 X_i；

End If2

Else If1 $p_i \geq 0.5$

8.采用螺旋游走策略更新 X_i；

End If1

End For

9.对 Ω_t 的 X_i 进行变异操作；

10.更新 Ω_t 中各 X_i 的适应值 $\{f_1(X_i), f_2(X_i)\}$；

11.令 $\Psi = \Gamma \cup \Omega$，采用多样性帕累托排序对 Ψ 中的解进行排序，更新非支配解集 *Leader_set*；

12.从 Ψ 选取前 N^p 个体添加到 Ω_{t+1}，令 $t = t+1$；

End While

13.采用理想点法从非支配解集 *Leader_set* 中选取并输出最佳调和解；

3.2 解的编码

本文采用实数编码法将特征子集 F_s 编码为群体中的解 X。假定原特征集包含 N 个特征，即 $F=\{f_1, f_2, \cdots, f_N\}$，则 $X=(x_1, x_2, \cdots, x_N)$ 为 $1\times N$ 的向量，其中 $x_i \in [-1,1], i=1,\cdots,N$。若 $x_i > 0$，则特征 f_i 被选择，若 $x_i \leqslant 0$，则特征 f_i 被剔除。

3.3 多样性帕累托排序

如何构建对群体中解的排序策略是群智能优化算法研究的关键内容。SPEA2[26] 是一个经典、高效的多目标遗传算法，其排序策略结合了帕累托法则和 K 近邻方法。该排序策略首先通过帕累托法则计算解的原始适应值(raw fitness) $R(X)$，再通过 K 近邻方法计算解在解空间的密度 $D(X)$，最后得到解的最终适应值 $F(X)=R(X)+D(X)$ 对解排序。$F(X) \in [0, N^2)$ 越小则解越优，若 $F(X)<1$ 则 X 为当前群体中的非支配解。然而，该策略用于特征选择会导致群体多样性在进化过程中迅速下降，影响算法优化性能。因为实数编码会出现不同解对应同一特征子集情况，从本质上来说这些解是重复的。例如，解(−0.3, 0.3, 0.2, −0.1, 0.5)和(−0.6, 0.7, 0.1, −0.3, 0.6)对应于同一特征子集 $\{f_2, f_3, f_5\}$。对于这些编码不同但实际相同的解，SPEA2 计算所得适应值是相等的，导致重复的解在进化过程中被保留，降低了群体的多样性。

为了解决以上问题，本文针对特征选择问题提出多样性帕累托排序方法，并将其用于 MWOA，具体步骤如算法 2。可以看到，所提排序方法首先将(解码后)非重复的解添加到集合 Ω_u，将冗余解添加到 Ω_r。接着，采用 SPEA2 的适应值计算法得到 Ω_u 中非重复解的适应值。对于 Ω_r 中的冗余解，其适应值为原适应值加惩罚变量 N^2。因此，Ω_r 中任意解的适应值都大于 Ω_u 中解的适应值。最后，根据适应值大小对解进行排序。综上，所提排序方法通过添加惩罚变量降低了冗余解的优先级，非重复的解在迭代过程中优先保留到下一代群体，保证了优化过程中群体的多样性。

算法 2 多样性帕累托排序

输入：群体 $\Omega = \{X_i \mid i=1,\cdots,N\}$

输出：排序后群体 Ω

/＊第一步：过滤冗余解＊/

1.令非重复解集合 $\Omega_u=\varnothing$，冗余解集为 $\Omega_r=\varnothing$；

2.将群体中第一个解 X_1 添加到 Ω_u；

For $i=1$ to N

3.如果 $decode(X_i)\in decode(\Omega_u)$，将 X_i 添加到 Ω_r，否则，将 X_i 添加到 Ω_u；

End For

/＊第二步：计算适应值并排序＊/

For each X in Ω_u

4.计算适应值 $F(X)=R(X)+D(X)$；

End For

For each X in Ω_r

5.从 Ω_u 查找与 X 解码后相同的解 X'，令 $F(X)=F(X')+N^2$；

End For

6.令 $\Omega=\Omega_u\cup\Omega_r$，并根据各解适应值 $F(X)$ 升序排列解，返回排序后的集合 Ω；

注：$decode(\cdot)$ 表示对解进行解码。

3.4　包围捕食策略

包围捕食模拟鲸鱼包围猎物的行为对鲸鱼位置（解）进行更新。假设第 t 次迭代群体中第 i 个解为 $X_i^t=(x_{i1}^t,x_{i2}^t,\cdots,x_{iN}^t)$，包围捕食策略将该解更新为

$$X_i^{t+1}=X^{*t}-A_i\cdot|C_i\,X^{*t}-X_i^t| \tag{6}$$

其中，X^{*t} 为当前迭代所得最优解。不同于单目标优化，多目标优化问题的最优解实际为非支配解集 $Leader\_set$ 中的多个解。为了解决该问题，随机从 $Leader\_set$ 中选取一个解作为式(6)中的 X^{*t} 计算 X_i^{t+1}。$A_i\cdot|C_i\cdot X_i^{t+1}-X_i^{t+1}|$ 为包围步长，$|\cdot|$ 表示对向量中的各元素依次求绝对值，A_i 和 C_i 定义如下：

$$A_i=2a_tr_i-a_t \tag{7}$$

$$C_i=2\cdot r_i' \tag{8}$$

其中，r_i 和 r_i' 为 $[0,1]$ 之间服从均匀分布的随机数；$a_t=2-2t/t_{\max}$ 为线性收敛因子，随着迭代次数增加由 2 均匀减小到 0，意味着鲸鱼对猎物的包围逐步收缩。

3.5　螺旋游走策略

螺旋游走模拟鲸鱼围绕猎物螺旋式游走的行为更新解。对于解 $X_i^t=(x_{i1}^t,x_{i2}^t,\cdots,x_{iN}^t)$，螺旋游走策略将该解更新为

$$X_i^{t+1}=D'e^{bl_i}\cos(2\pi l_i)+X_i^t \tag{9}$$

其中 $D'=|X^{*t}-X_i^t|$ 表示当前鲸鱼位置 X_i^t 与猎物位置 X^{*t} 的距离，同理，X^{*t} 为从 $Leader\_set$ 中随机选择的解。b 为人工设置参数，参考文献[18]，本文选取 $b=1$。l_i 为 $[-1,1]$ 之间服从均匀分布的随机数。

实际上鲸群采取的是边螺旋游走、边包围猎物的捕食行为，这一行为称为泡泡网觅食（bubble-net attacking）。为了模拟这一行为，每个鲸鱼以 0.5 的概率随机选择螺旋游走或包围捕食策略。

3.6　猎物搜寻策略

鲸鱼会根据群体其他鲸鱼的位置采取一种相对随机的猎物搜寻策略。鲸鱼 i 采取包围捕食或猎物搜寻策略决定于 A_i 的值。如果 $|A_i|<1$ 则解 X_i^t 采用式(6)的包围捕食策略更新得到 X_i^{t+1}。如果 $|A_i|\geq1$，则 X_i^t 采取如下猎物搜寻策略更新解：

$$X_i^{t+1}=X_{rand}^t-A_i\cdot|C_i\,X_{rand}^t-X_i^t| \tag{10}$$

其中 X_{rand}^t 为从当前群体中随机选择的一个解，A_i 和 C_i 的计算方法如式(7)和(8)。可以看到，猎物搜寻与包围捕食策略的主要区别在于猎物搜寻选择当前群体中的随机解为包围中心。猎物搜寻策略的是为了增强算法探索（广域搜索）能力而提出的，然而，该策略不能解决算法在中后期探索能力下降缺点。因为，随机解 X_{rand}^t 的选择是从群体中选择的，在算法中后期群体中的解整体向趋向于一致。因而，需要在优化策略中引入其他策略，进一步增强算法探索能力。本文在 MWOA 中引入变异操作增强算法探索能力。

3.7　变异操作

在采用包围捕食策略、螺旋游走策略或猎物搜寻策略更新解之后，对群体中的解进行变异操作。本文针对实数编码的解提出如下变异方法，若 $X=(x_1,x_2,\cdots,x_N)$ 为群体中的解，经过变异后得到解 $X'=(x_1',x_2',\cdots,x_N')$，

$$x_i'=\begin{cases}x_i, i\neq m_p\\ -x_i, i=m_p\end{cases};i=1,\cdots,N \tag{11}$$

其中 $m_p\in\{1,2,\cdots,N\}$ 为随机选择的变异点。为了平衡算法探索能力和开发能力，每次迭代只随机选取群体中一半的解进行变异操作。

4　算例分析

4.1　实验设置

为了验证算法进行 KQC 识别的有效性，本文选取 3 组制造过程数据进行实验分析。分别是胶乳产品制造过程数据 LATEX、尼龙产品制造过程数据 ADPN 和纸质产品制造过程数据 PAPER[27,28]。文献[6]和[9]在原数据基础上，根据质量性能变量 y 的值，将产品质量水平划分为合格\不合格，并在此基础上进行产品 KQC 识别。由于

数据集中合格产品是多于不合格产品的，这三组数据都是非平衡的，具体信息如表 2。

表 2 数据集信息

数据集	样本数	合格品	不合格品	质量特性数
LATEX	262	184	78	117
ADPN	71	51	20	100
PAPER	384	351	33	54

本文选取 SFS、SBS、MCP、WOA－CM、NSGAII－IPM、CMDPSOFS、NSPSOFS 等特征选择算法作为对比方法。SFS 和 SBS[11] 为基于爬山策略的经典特征选择算法。MCP[5] 基于偏最小二乘回归模型构建质量特性重要性指标识别 KQC。WOA－CM[23] 为最近提出的基于单目标 WOA 的特征选择算法，该方法在 WOA 中嵌入了遗传算法常用的交叉、变异操作，优化目标综合了最小化分类错误率和特征数。NSGAII－IPM、CMDPSOFS、NSPSOFS 为基于多目标优化模型的特征选择算法。NSGAII－IPM[9] 采用了最小化错误率、第 II 类错误率和特征数的多目标优化模型，并用改进的经典多目标遗传算法 NSGA－II 和理想点法选择最优 KQC 集。CMDPSOFS 和 NSPSOFS[10] 为最近提出的基于多目标粒子群优化的特征选择算法，优化目标为最小化错误率和特征数。在 CMDPSOFS 和 NSPSOFS 中，同样使用理想点法从多个解中选择最优 KQC 集。MCP、NSGAII－IPM、WOA－CM、CMDPSOFS、NSPSOFS 基于 MATLAB R2012b 实现，SFS、SBS 直接使用机器学习工具 Weka 3.7[29] 实现。由于 NSGAII－IPM、WOA－CM、CMDPSOFS、NSPSOFS 等都基于群体进化，为了保证实验公平性，这些算法选择相同的迭代次数 $t_{\max} = 200$ 和相同的群体规模 $N^p = 100$。MCP、NSGAII－IPM、WOAFS、CMDPSOFS、NSPSOFS 的其他参数设置与文献中保持一致，SFS 和 SBS 采用 Weka 中默认设置。

实验采用分层 10 折交叉验证法。该方法首先将数据划分为 10 部分，依次选择其中 9 部分为训练集进行 KQC 识别和训练分类器，另 1 部分为测试集验证 KQC 识别的有效性，整个实验重复 10 次。在算法性能对比中采用 10 次结果的平均值。采用两方面性能指标验证 KQC 识别的有效性，包括所识别 KQC 集对测试集产品质量水平（是否合格）的分类能力和识别的 KQC 数。若算法识别 KQC 集分类能力高且识别 KQC 数少，则说明算法过滤无关质量特性的能力强，KQC 识别的效果好。由于数据为非平衡数据，具体采用分类精度、敏感性和特异性三个指标度量分类能力。另外，实验中的分类器选取简单、高效的朴素贝叶斯分类器（naïve Bayes classifier）[30]，该分类器调用自 Weka 3.7，并采用默认设置。所有实验均在一台 Intel Core i5－3470 3.2GHz CPU 和 4G 内存的计算机实现。

4.2 实验结果与分析

表 3 所示为各算法所得分类精度。可以看到，在 LATEX 和 ADPN 数据集，本文算法得到最高的分类精度，分别为 81.92％和 85.89％。在 PAPER 数据集，WOA－CM 所得分类精度最高，达到 90.35％，本文算法所得分类精度略低，为 88.02％。为了便于比较各算法分类精度结果，表 3 同时给出各算法在三个数据集的平均分类精度。可以看到本，本文算法在三个数据集的平均分类精度最高，为 85.28％，其次为另外三个基于多目标优化的特征选择算法（即 NSGAII－IPM、CMDPSOFS、NSPSOFS）、基于单目标优化的 WOA－CM 和基于偏最小二乘模型的 MCP，而传统特征选择算法 SFS 和 SBS 表现最差。总体上，本文算法在各数据集都获得较高的分类精度，表明了算法所识别 KQC 集对产品质量水平有不错预测能力。

表 3 各算法分类精度（％）对比

	本文算法	SFS	SBS	MCP	WOA－CM	NSGAII－IPM	CMDPSOFS	NSPSOFS
LATEX	81.92	78.62	76.75	81.31	77.12	81.71	80.16	75.97
ADPN	85.89	77.32	75.71	79.11	81.61	83.99	79.79	80.83
PAPER	88.02	82.06	87.99	89.82	90.35	88.54	90.09	89.99
平均	85.28	79.33	80.15	83.41	83.03	84.75	83.35	82.26

由于实验所用制造过程数据集为非平衡数据，仅用分类精度并不能准确度量各算法所识别 KQC 集对产品质量水平的预测能力。为了准确度量各算法对正负（不合格\合格）类样本的分类效果，表 4 给出了各算法所得敏感性

和特异性精度。其中，敏感性反映对正类样本的分类效果，特异性反映对负类样本的分类效果。根据敏感性结果，在 ADPN 和 PAPER 数据集，本文算法都取得最高精度，分别是 90%和 90.83%，NSGAII－IPM 所得精度略低于本文算法，分别是 85%和 82.50%，而其他算法所得敏感性精度明显低于这两个算法。在 LATEX 数据集，MCP 取得最高的 83.57%的敏感性精度，其次为本文算法 74.46%和 NSGAII－IPM 73.39%的敏感性精度，其他算法所得敏感性精度远低于这三个算法。根据平均结果，仅本文算法和 NSGAII－IPM 敏感性精度达到 80%以上，WOA－CM 达到 70%以上，其他算法均在 60%多。根据特异性结果，在 LATEX、ADPN 和 PAPER 数据集，取得最高精度的分别是 SFS(90.70%)、MCP(88.33%)和 NSP-SOFS(92.32%)，本文算法所得精度略低，分别是85.06%、84.33%和 87.75%。综合敏感性和特异性的结果，MCP 在 LATEX 数据集同时取得较高敏感性和特异性精度，但在其他数据集所得敏感性精度明显低于特异性精度，SFS、NSPSOFS 和 CMDPSOFS 等算法总体取得很高的特异性精度，但是所得敏感性精度远低于特异性精度，说明这些方法所识别 KQC 集仅对数据集中占多数的负类(合格)样本有高的分类精度，而对占少数的正类(不合格)样本分类精度较差，数据非平衡对算法的 KQC 识别造成了影响，所识别 KQC 是有偏差的。对于本文算法和 NSGAII－IPM，在各数据集所得敏感性和特异性精度都保持在较高水平，表明了处理数据非平衡性的有效性。

表 4　各算法敏感性\特异性(%)对比

	本文算法	SFS	SBS	MCP	WOA－CM	NSGAII－IPM	CMDPSOFS	NSPSOFS
敏感性								
LATEX	74.46	49.82	63.93	83.57	62.50	73.39	63.48	52.68
ADPN	90.00	75.00	65.00	55.00	80.00	85.00	60.00	63.33
PAPER	90.83	58.33	80.83	62.50	76.67	82.50	68.33	67.50
平均	85.10	61.05	69.92	67.02	73.06	80.30	63.94	61.17
特异性								
LATEX	85.06	90.70	82.13	80.47	83.16	85.32	87.18	85.88
ADPN	84.33	78.00	80.00	88.33	82.00	83.67	87.33	87.67
PAPER	87.75	84.36	88.88	92.31	91.73	89.17	92.02	92.32
平均	85.72	84.35	83.67	87.04	85.63	86.05	88.85	88.62

表 5 所示为各算法所识别 KQC 数，表中括号内列出了所识别 KQC 数占总质量特性数的百分比。可以看到，本文算法在各数据集都识别出最少的 KQC，说明了该算法能够有效进行特征降维；SFS、NSGAII－IPM、CMDP-SOFS 和 NSPSOFS 所识别 KQC 数略多于本文算法；SBS、WOA－CM 和 MCP 等算法在各数据集识别 KQC 都明显多于其他算法(除了 MCP 在 ADPN 数据集识别较少 KQC)。另外可以看到，降维能力较强的 NSGAII－IPM、CMDPSOFS、NSPSOFS 和本文算法都基于多目标特征选择模型。说明在建立特征选择模型时，把最小化特征数单独作为一个优化目标能够取得不错特征降维效果。WOA－CM 虽然在建立模型时考虑到了特征数这一因素，但是由实验结果可以看到基于单目标优化的 WOA－CM 进行特征降维的有效性明显不如多目标优化方法。另外，从实验结果来看，虽然 SFS 和 SBS 同样基于爬山策略，但基于正向选择的 SFS 有更好的特征降维能力。

表 5　各算法识别 KQC 数

	本文算法	SFS	SBS	MCP	WOA－CM	NSGAII－IPM	CMDPSOFS	NSPSOFS
LATEX	3.8(3.2%)	7.7(6.6%)	93.8(80.2%)	27.4(23.4%)	52.1(44.5%)	5.8(5.0%)	5.7(4.9%)	5.4(4.6%)
ADPN	2.3(2.3%)	5.3(5.3%)	25.8(25.8%)	2.8(2.8%)	27.0(27.0%)	2.5(2.5%)	3.7(3.7%)	4.0(4.0%)

PAPER	3.0(5.6%)	4.0(7.4%)	18.2(33.7%)	48.9(90.6%)	25.9(48.0%)	3.8(7.0%)	4.5(8.3%)	6.1(11.3%)
平均	3.0(3.7%)	5.7(6.4%)	45.9(46.6%)	26.4(38.9%)	35.0(39.8%)	4.0(4.8%)	4.6(5.6%)	5.2(6.6%)

综合表 3、4、5 的结果，本文算法不仅能够识别最少 KQC，同时所识别 KQC 集对不同质量产品都有高水平分类能力。由表 4 结果可以看到，仅本文算法和 NSGAII－IPM 在各数据集能够同时获得较高敏感性精度和特异性精度，其他算法总体上所得敏感性精度要远低于特异性精度。造成这一结果的原因是，其他算法没有考虑到数据非平衡对 KQC 识别所造成的影响，造成所识别 KQC 对少量不合格品（正类样本）的分类效果不佳。而本文算法和 NSGAII－IPM，分别应用了 G－mean 和第二类错误率，使得算法在识别 KQC 时充分考虑到对合格品的分类效果。另外，对比本文算法和 NSGAII－IPM，本文算法总体上得到更高得分类精度，同时识别得到更少 KQC，表明了本文算法能够更为有效识别 KQC。综合来看，针对非平衡制造过程数据，本文算法取得了最佳 KQC 识别结果。

为了进一步对比算法性能，表 6 给出了各算法 CPU 运行时间。可以看到，MCP 在各数据的 CPU 运行时间明显低于其他算法，其次为 SFS。造成这一结果的原因是，MCP 算法仅依靠偏最小二乘模型对各质量特性进行权重排列，没有复杂的优化过程。但是可以从表 3 至 5 看到该算法在 ADPN 和 PAPER 数据集的 KQC 识别效果较差，并且进行特征降维的能力也较差。SFS 的 CPU 运行时间较低的原因是该算法采用了爬山策略，因而优化迭代次数远低于其他采用进化、群智能优化策略的算法。但是，同样采用爬山策略的 SBS 在各数据集的 CPU 运行时间远高于 SFS，甚至在 LATEX 和 ADPN 数据集高于采用多目标优化策略的 NSGAII－IPM、CMDPSOFS、NSPSOFS 和本文算法。造成这一结果的原因是：特征子集越大（包含特征越多），适应值估计所需时间越高；SBS 从一个完整的特征集合开始逐步过滤特征，并且特征过滤效率较低，导致该算法估计的特征子集远大于其他算法，耗费了大量时间。基于单目标优化的 WOA－CM 和基于多目标优化的特征选择算法采用了相同的群体规模和迭代次数，理论上这些算法估计特征子集适应值的次数是一致的。但是，由实验结果可以看到 WOA－CM 耗时明显高于基于多目标优化的特征选择算法。原因与 SBS 类似，WOA－CM 降低特征维度的效率较低，导致估计特征子集有效性所耗时间更多。本文算法与 NSGAII－IPM、CMDPSOFS、NSPSOFS 相比，除了在 LATEX 数据集运行时间略高于 NSGAII－IPM，在其他数据集均耗时最短。总体上，本文所提 MWOA 在时间复杂度上与传统基于遗传算法和粒子群优化的多目标优化算法接近。

表 6 各算法 CPU 运行时间(S)

	本文算法	SFS	SBS	MCP	WOA－CM	NSGAII－IPM	CMDPSOFS	NSPSOFS
LATEX	1.47E+04	9.86E+02	4.41E+04	3.38E+02	5.07E+04	1.31E+04	1.83E+04	1.72E+04
ADPN	2.70E+03	1.91E+02	8.88E+03	3.70E+01	6.03E+03	3.09E+03	3.45E+03	3.34E+03
PAPER	1.08E+04	1.93E+02	1.02E+04	1.10E+02	3.59E+04	1.45E+04	1.76E+04	1.89E+04

5 结论

本文提出基于 MWOA 的多目标特征选择算法用于非平衡制造过程数据进行 KQC 识别。针对数据非平衡性，建立了最大化 G－mean 和最小化特征数的多目标特征选择模型。为了求解该模型，基于 WOA 算法中的包围捕食策略、螺旋游走策略和猎物搜寻策略，提出 WOA 的多目标优化版本——MWOA。所提 MWOA 采用了一种针对特征选择问题的多样性帕累托排序方法，并嵌入了变异操作，保证算法优化性能。选取 3 组非平衡制造过程数据对算法有效性进行了验证。实验结果表明，所提算法能够在各数据集得到高水平分类性能的同时明显降低识别 KQC 数，表明了算法进行 KQC 识别的有效性。另外，所提 MWOA 的运行时间与传统的多目标遗传和粒子群优化算法接近。针对产品质量水平为多个分级的非平衡制造过程数据提出 KQC 识别方法，改进 KQC 识别模型以降低算法时间复杂度，是未来研究的两个方向。

参考文献：

[1] Lee D J, Thornton A C. The identification and use of key characteristics in the product development process [C]. 1996 ASME Design Engineering Tech-

nical Conference. 1996.

[2] 吴锋，马义中，朱连燕. 基于 Kriging 模型的复杂产品制造过程稳健参数设计与控制 [J]. 系统工程，2014，32(07)：81－86.

[3] Li A-D，He Z，Wang Q，et al. Key quality characteristics selection for imbalanced production data using a two-phase bi-objective feature selection method [J]. European Journal of Operational Research，2019，274(3)：978－989.

[4] Pierre E S，Tuv E. Robust，non-redundant feature selection for yield analysis in semiconductor manufacturing [C]. Industrial Conference on Data Mining. Springer Berlin Heidelberg，2011：204－217.

[5] Anzanello M J，Albin S L，Chaovalitwongse W A. Selecting the best variables for classifying production batches into two quality levels [J]. Chemometrics and Intelligent Laboratory Systems，2009，97(2)：111－117.

[6] Anzanello M J，Albin S L，Chaovalitwongse W A. Multicriteria variable selection for classification of production batches [J]. European Journal of Operational Research，2012，218(1)：97－105.

[7] 王化强，牛占文. 基于 LASSO 的复杂产品关键质量特性识别 [J]. 系统工程，2014，32(06)：137－141.

[8] Li A-D，He Z，Zhang Y. Bi-objective variable selection for key quality characteristics selection based on a modified NSGA-II and the ideal point method [J]. Computers in Industry，2016，82：95－103.

[9] 李岸达，何桢，何曙光. 基于 NSGA-II 的非平衡制造数据关键质量特性识别 [J]. 系统工程理论与实践，2016，36(6)：1472－1479.

[10] Xue B，Zhang M，Browne W N. Particle swarm optimization for feature selection in classification：a multi-objective approach [J]. IEEE transactions on cybernetics，2013，43(6)：1656－1671.

[11] Kohavi R，John G H. Wrappers for feature subset selection [J]. Artificial intelligence，1997，97(1)：273－324.

[12] Unler A，Murat A. A discrete particle swarm optimization method for feature selection in binary classification problems [J]. European Journal of Operational Research，2010，206(3)：528－539.

[13] Ahila R，Sadasivam V，Manimala K. An integrated PSO for parameter determination and feature selection of ELM and its application in classification of power system disturbances [J]. Applied Soft Computing，2015，32：23－37.

[14] 李岸达，何桢，何曙光. 基于 GSA 的复杂产品关键质量特性识别 [J]. 系统工程与电子技术，2015，37(9)：2073－2079.

[15] Mistry K，Zhang L，Neoh S C，et al. A Micro-GA Embedded PSO Feature Selection Approach to Intelligent Facial Emotion Recognition [J]. IEEE Transactions on Cybernetics，2017，47(6)：1496－1509.

[16] De Stefano C，Fontanella F，Marrocco C，et al. A GA-based feature selection approach with an application to handwritten character recognition [J]. Pattern Recognition Letters，2014，35：130－141.

[17] Soyel H，Tekguc U，Demirel H. Application of NSGA-II to feature selection for facial expression recognition [J]. Computers & Electrical Engineering，2011，37(6)：1232－1240.

[18] Mirjalili S，Lewis A. The Whale Optimization Algorithm [J]. Advances in Engineering Software，2016，95：51－67.

[19] 龙文，蔡绍洪，焦建军，等. 求解大规模优化问题的改进鲸鱼优化算法 [J]. 系统工程理论与实践，2017，37(11)：2983－2994.

[20] Aljarah I，Faris H，Mirjalili S. Optimizing connection weights in neural networks using the whale optimization algorithm [J]. Soft Computing，2016：1－15.

[21] Nazari-Heris M，Mehdinejad M，Mohammadi-Ivatloo B，et al. Combined heat and power economic dispatch problem solution by implementation of whale optimization method [J]. Neural Computing & Applications，2017：1－16.

[22] Mafarja M M，Mirjalili S. Hybrid Whale Optimization Algorithm with simulated annealing for feature selection [J]. Neurocomputing，2017，260：302－312.

[23] Mafarja M M，Mirjalili S. Whale Optimization Approaches for Wrapper Feature Selection [J]. Applied Soft Computing，2018，62：441－453.

[24] Sharawi M, Zawbaa H M, Emary E, et al. Feature selection approach based on whale optimization algorithm [C]. Ninth International Conference on Advanced Computational Intelligence. IEEE, 2017: 163－168.

[25] 杨光飞，崔雪娇，张翔. 基于抽样和规则的不平衡数据关联分类方法 [J]. 系统工程理论与实践，2017，37(4)：1035－1045.

[26] Ziztler E, Laumanns M, Thiele L. SPEA2: Improving the strength Pareto evolutionary algorithm for multiobjective optimization [J]. Evolutionary Methods for Design, Optimization, and Control, 2002: 95－100.

[27] Gauchi J P, Chagnon P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data [J]. Chemometrics & Intelligent Laboratory Systems, 2001, 58(2): 171－193.

[28] Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics [J]. Chemometrics and intelligent laboratory systems, 2001, 58(2): 109－130.

[29] Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update [J]. ACM SIGKDD explorations newsletter, 2009, 11(1): 10－18.

[30] John G H, Langley P. Estimating continuous distributions in Bayesian classifiers [C]. Proceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1995: 338－345.

A Multi-Objective Whale Optimization Algorithm for Key Quality Characteristics Identification

LI An-da[1], HE Zhen[2], WANG Qing[1]

(1. School of Management, Tianjin University of Commerce, Tianjin 300134, China;
2. College of Management and Economics, Tianjin University, Tianjin 300072, China)

Abstract: In this paper, a feature selection method based on a multi-objective whale optimization algorithm (MWOA) is proposed for the key quality characteristics (KQC) identification problem. First, the identification model is established as maximizing the G-mean metric and minimizing feature (quality) subset size for imbalanced manufacturing data. Second, a swarm-based optimization algorithm, i.e., MWOA, is proposed to solve this model. MWOA adopts a new diversity Pareto sorting strategy, which sorts solutions in the swarm as well as keeps the swarm diversity for feature selection problems. Moreover, a mutation operator is embedded in MWOA as the search strategy of whale optimization can easily get trapped in the local optimum. The experimental results on 3 imbalanced manufacturing datasets illustrate that the proposed method can effectively identify KQCs.

Key words: key Quality Characteristics Identification; Feature Selection; Multi-Objective Optimization; Whale Optimization Algorithm; Imbalanced Data