doi: 10.12011/1000-6788(2016)06-1472-08 中图分类号: F273.2 文献标志码: A

基于 NSGA-II 的非平衡制造数据关键质量特性识别

李岸达, 何 桢, 何曙光

(天津大学 管理与经济学部, 天津 300072)

摘 要 针对非平衡产品制造数据关键质量特性 (critical to quality characteristics, CTQs) 识别, 提出基于 NSGA-II 的特征选择算法. 首先, 在分类错误率和特征子集大小基础上, 针对数据非平衡性, 引入第 II 类错误率度量质量特性子集的重要性. 接着, 应用多目标进化算法 NSGA-II 最小化以上三个度量标准, 得到非支配解集. 最后, 引入理想点法从非支配解集中选择最佳调和解, 得到 CTQ 集. 算例结果表明, 所提算法能够得到较高分类精度, 同时有效降低第 II 类错误率与 CTQ 集大小, 说明了算法的有效性.

关键词 非平衡数据; 特征选择; 关键质量特性; NSGA-II; 理想点法; 第 II 类错误率

Critical to quality characteristics identification for imbalanced production data based on NSGA-II

LI Anda, HE Zhen, HE Shuguang

(College of Management and Economics, Tianjin University, Tianjin 300072, China)

Abstract To select critical to quality characteristics (CTQs) for imbalanced production data, a feature selection algorithm based on NSGA-II is proposed. Firstly, to solve the problem of data imbalance, type II error is introduced to measure the importance of quality characteristics subset in addition to classification error and feature subset size. Secondly, NSGA-II, a multi-objective evolutionary algorithm, is applied to minimize the three metrics above, and a non-dominated solution set is acquired. Finally, the ideal point method is adopted to obtain the best compromise solution (CTQ set) from the non-dominated solution set. Experimental results illustrate that the proposed algorithm can obtain high classification accuracy, and in the meantime, effectively reduce type II error and CTQ set size, which shows the efficiency of the proposed algorithm.

Keywords imbalanced data; feature selection; critical to quality characteristics (CTQs); NSGA-II; the ideal point method; type II error

1 引言

现代化工业生产中, 有大量产品质量特性 (包括产品特性, 过程变量, 装配特性等 [1]) 从生产制造过程中被收集. 然而, 这些质量特性并不同等重要, 有些质量特性对产品质量有重要影响, 有些对产品质量影响甚微, 因此从大量质量特性中识别影响产品质量的关键质量特性 (critical to quality characteristics, CTQs) 是一项重要任务. 通过识别产品 CTQ, 可以有效控制产品制造成本, 提高产品质量 [2]. CTQ 识别可以看作是

收稿日期: 2015-01-15

作者简介: 李岸达 (1989–), 男, 汉, 甘肃渭源人, 博士研究生, 研究方向: 质量工程、智能算法, E-mail: andali1989@163.com; 何桢 (1967–), 男, 汉, 河南台前人, 教授, 博士生导师, 博士, 研究方向: 质量工程与质量管理、工业工程, E-mail: zhhe@tju.edu.cn; 何曙光 (1975–), 男, 汉, 内蒙古呼和浩特人, 教授, 博士, 研究方向: 质量工程, E-mail: shuguanghe@126.com.

基金项目: 国家杰出青年科学基金 (71225006); 国家自然科学基金 (71102140)
Foundation item: National Natural Science Funds for Dstinguished Young Scholars (71225006); National Natural Science Foundation of China (71102140)
中文引用格式: 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.
英文引用格式: Li A D, He Z, He S G. Critical to quality characteristics identification for imbalanced production data based on NSGA-II[J]. Systems Engineering — Theory & Practice, 2016, 36(6): 1472–1479.

一个特征选择 (feature selection) 问题 [2−4]. 特征选择将影响样本类别 (产品质量水平) 的重要特征 (质量特性) 选择出来, 并剔除冗余、无关特征, 因此该类方法能够有效识别产品 CTQ.

通常, 根据特征子集有效性的评价标准, 特征选择可以划分为: Filter 算法和 Wrapper 算法. Filter 算法在应用学习算法之前通过不同特征度量标准 (距离、信息、一致性等) 评估特征子集与类标签的相关性, 进而选择关键特征, 该过程是独立于学习算法的一个预处理过程 [5]. 常见的 Filter 算法包括 FCBF、ReliefF 等 [5,6]. Filter 算法的优点是算法的时间复杂度较低. Wrapper 算法对特征子集进行评估时引入了学习算法, 学习算法的分类性能是评价特征子集有效性的重要度量 [7]. 常见的 Wrapper 算法包括 SFS、SBS 等 [8]. 相对于 Filter 算法, Wrapper 算法时间复杂度较高, 但该类算法通常拥有更高的分类精度.

特征选择可以看作是一个多目标优化问题, 其目标通常是最大化特征子集的分类精度以及最小化特征子集大小 (所含特征数)[9]. 因此, 搜索策略是特征选择的一个重要部分. 根据搜索策略, 特征选择可以划分为确定性算法 (如 SFS、SBS [8]) 和随机算法 (如粒子群优化 (PSO)、遗传算法 (GA) [10−14]). 文献 [12] 在应用 PSO 算法进行特征选择时, 将剔除特征数占总特征数的比引入适应度函数, 建立基于特征数和分类精度的综合适应度函数进行优化. 文献 [13] 将每个特征所占成本引入 GA 的适应度函数, 建立基于分类精度与成本的综合适应度函数, 应用于 SVM 分类器的特征选择问题. 以上方法通过建立综合适应度函数, 将多目标优化问题转变为单目标问题, 并使用优化算法进行优化. 但是, 在缺少相关领域知识的情况下, 在建立综合适应度函数时很难设定各子目标的权重. 随着多目标优化的发展, 多目标进化算法 (如 MOGA、NSGA-II、MOPSO[15−17]) 被引入特征选择领域, 该类算法能够同时对多个目标进行优化, 有效避免了不易确立综合适应度函数的难点.

通过对分类精度以及特征子集大小进行多目标优化, 能够在得到不错分类精度的同时有效控制特征子集的大小. 然而, 对于非平衡数据, 仅得到高的分类精度并不能说明特征子集的有效性 [2]. 例如, 数据集中有 10% 正 (少) 类样本, 90% 负 (多) 类样本, 如果学习算法将所有样本都判定为负类样本, 其分类精度达到 90%, 但是第 II 类错误率 (正类样本错分率) 为 100%, 显然在这种情况下并不能仅以分类精度为标准来判断特征子集的重要性. 在实际生产制造中, 产品合格率通常较高, 合格产品数量多于非合格产品, 从生产线收集而来的产品质量特性数据集是一个非平衡数据, 所以产品关键质量特性识别问题通常是一个非平衡数据特征选择问题. 对于非平衡数据, 第 II 类错误率可以度量特征子集对少类样本的分类性能, 因此应当引入第 II 类错误率作为特征子集有效性的度量标准 [18].

为了解决非平衡产品数据集的 CTQ 识别问题, 本文提出基于 Wrapper 框架的特征选择算法 —— 基于 NSGA-II 的非平衡数据特征选择算法. 首先, 确立优化目标为最小化分类错误率、第 II 类错误率以及特征子集大小, 并应用多目标进化算法 NSGA-II[19] 优化多个目标, 得到一个非支配解集. 接着, 引入理想点法, 从非支配解集中选择出最佳调和解, 即 CTQ 集. 最后, 将本文算法应用于 3 组非平衡制造数据的 CTQ 识别, 验证了算法的有效性.

2 产品 CTQ 识别框架

产品 CTQ 识别问题可以定义为: 令 D 为产品数据集, $Q = \{q_1, q_2, \cdots, q_N\}$ 为产品的质量特性 (特征) 集; $C \in \{1, -1\}$ 是类标签, 代表产品的质量水平, 其中 "1" 和 "−1" 分别代表 "不合格"、"合格". 数据集包含 M 个样本 $\boldsymbol{X_i} = (x_{i1}, x_{i2}, \cdots, x_{iN}, y_i)$ $(i = 1, 2, \cdots, M)$, 其中 $x_{i1}, x_{i2}, \cdots, x_{iN}$ 为该样本各质量特性 $q_1, q_2, \cdots, q_N$ 的观测值; y_i 是样本质量水平 (类标签) C 的观测值. CTQ 识别的目的是从 Q 中挑选出影响产品质量水平 C 的质量特性子集作为产品 CTQ 集.

CTQ 集的有效性可以用分类精度 (或错误率)、第 II 类错误率以及 CTQ 集大小三个标准来度量. 分类精度反映了 CTQ 集对产品质量水平的预测能力. CTQ 集大小反映了 CTQ 集所含质量特性数的多少. CTQ 集越小, 说明识别算法能够更有效过滤无关、冗余质量特性. 在实际产品生产制造中, 产品制造数据集通常是非平衡的, 分类精度更多地受到负类样本 (合格产品) 的影响. 第 II 类错误率度量了对正类样本 (不合格产品) 的错分程度. 因此, 引入第 II 类错误率 (见 3.1 节) 来衡量 CTQ 集对正类样本的预测能力十分必要. 综上, 应用特征选择算法优化质量特性子集的分类精度、第 II 类错误率以及子集大小, 可以从原始质量特性中识别出产品 CTQ.

为了验证 CTQ 识别的有效性, 需要一部分未参与 CTQ 识别的数据作为测试集, 因此在 CTQ 识别之前要将数据集划分为训练集和测试集. 训练集用来输入特征选择算法进行 CTQ 识别并得到 CTQ 集. 识别

所得 CTQ 集和训练集能够建立学习算法, 用该学习算法可以对测试集数据的质量水平进行预测得到分类结果. 通过测试集分类结果可以验证所选 CTQ 集的有效性.

基于以上分析, 本节提出产品 CTQ 识别框架, 如图 1 所示. 识别框架分 5 步: 1) 将数据集划分为训练集与测试集. 2) 输入训练集到特征选择算法进行 CTQ 识别, 得到 CTQ 集. 3) 应用 CTQ 集与训练集建立学习算法. 4) 预测测试集产品的质量水平, 得到分类结果. 5) 评估 CTQ 识别效果.

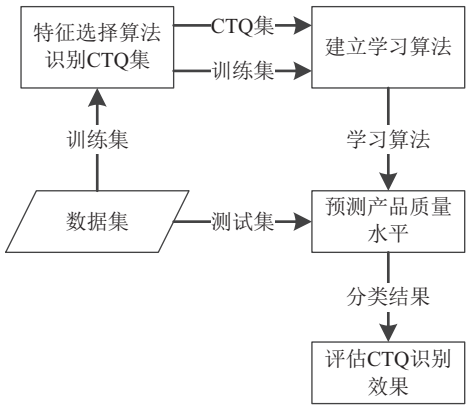

图 1 产品 CTQ 识别框架

3 建立特征选择算法

本节建立非平衡数据特征选择算法. 根据第 2 节分析, 算法需要搜索特征子集以优化三个目标: 1) 最小化特征子集的分类错误率; 2) 最小化特征子集的第 II 类错误率; 3) 最小化特征子集大小. 由于在估计分类错误率与第 II 类错误率时需要引入学习算法, 因此所提算法是基于 Wrapper 框架的. 搜索算法选用一个经典、高效的多目标进化算法 NSGA-II[19]. 由于 NSGA-II 返回的非支配解集通常包含多个解 (特征子集), 因此本文引入理想点法 [20], 从非支配解集中选择最佳调和解作为最优特征子集. 综上, 本文算法分两阶段实现, 首先应用 NSGA-II 优化目标函数得到非支配解集, 接着应用理想点法从非支配集中选择最佳调和解, 流程如图 2 所示.

图 2 特征选择算法流程

3.1 目标函数

本文特征选择算法优化三个目标函数, 分别是: 分类错误率、第 II 类错误率以及特征子集大小. 分类错误率度量学习算法对所有样本错分的程度, 第 II 类错误率度量学习算法对正类样本错分的程度, 特征子集大小表示特征子集所含特征数. 三个目标函数的计算方法介绍如下.

表 1 所示为混淆矩阵, TP 表示学习算法将正类样本分类为正类的数量, FP 表示学习算法将负类样本分类为正类的数量, FN 表示学习算法将正类样本分类为负类的数量. TN 表示学习算法将负类样本分类为负类的数量. 则分类错误率和第 II 类错误率计算公式分别如式 (1)、式 (2) 所示.

表 1 混淆矩阵

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

$$Error = \frac{FN + FP}{TP + FN + FP + TN} \tag{1}$$

$$TypeII = \frac{FN}{TP + FN} \tag{2}$$

本文使用训练集内部 k 折交叉验证对特征子集的分类错误率和第 II 类错误率进行估计, 该方法是 Wrapper 算法的常用估计方法 [7]. 步骤如下: 1) 将训练集分为 k 折. 2) 选取第 r ($r = 1, 2, \cdots, k$) 折作为内部测

试集, 其他 $k-1$ 折作为内部训练集, 得到第 r 折的分类错误率与第 II 类错误率. 3) 平均 k 折的结果, 得到特征子集分类错误率与第 II 类错误率的估计. 参照文献 [7], 本文选取 $k=5$. 第三个目标函数特征子集大小, 可以通过计算特征子集中所含特征个数得到.

3.2 NSGA-II 算法

NSGA-II 算法由 Deb 提出, 是多目标进化算法 NSGA [21] 的改进. 相对于 NSGA, NSGA-II 的时间复杂度更小, 进化过程能够保留每一代的精英个体, 不需要单独为算法的群体多样性策略设置实验参数. 因此, 本文选取 NSGA-II 作为特征选择的搜索算法.

3.2.1 算法步骤

本文特征选择问题是一个多目标优化问题, 定义如式 (3). 其中 $F(s)=\{f_1(s),f_2(s),f_3(s)\}$ 为目标函数, f_1 为分类错误率, f_2 为第 II 类错误率, f_3 为特征子集大小; s 代表解, 即特征子集; ϕ 代表可行解集. NSGA-II 通过优化多目标 $F(s)$ 最终返回一个非支配解集 ψ, ψ 中的每一个解代表一个特征子集.

$$\begin{aligned} &\min \quad F(s)=\{f_1(s),f_2(s),f_3(s)\} \\ &\text{s.t.} \quad s\in\phi \end{aligned} \tag{3}$$

表 2 所示为基于 NSGA-II 的特征选择步骤. 表中 "快速非支配排序" (fast non-dominated sorting) 是一个非支配等级排序方法, 它将群体中的解根据 Pareto 支配关系进行分级, 各解被赋予不同非支配等级, 相同等级的解在相同的非支配前沿面上; "拥挤距离分配" (crowding distance assignment) 根据每个解与相同前沿面相邻解的距离计算得到该解的拥挤距离, 拥挤距离体现了解在前沿面上的拥挤程度; "拥挤对比算子" (crowded comparison operator) 对比两个解的优劣, 非支配等级低的解优于非支配等级高的解, 若两个解在相同前沿面上, 则拥挤距离更小的解更优; 详见文献 [19]. 算法编码方法以及遗传算子 (包括选择方法, 交叉方法, 变异方法) 见 3.2.2 节与 3.2.3 节.

表 2　基于 NSGA-II 的特征选择

输入: 训练集 (含 N 个特征)
输出: 非支配解集 ψ
1. 令 $t=0$.
2. 随机产生初始群体 P_0, 包含 N_p 个体 (可行解).
3. 评价 P_0 中每个个体 s 的多目标 $F(s)=\{f_1(s),f_2(s),f_3(s)\}$.
4. 使用快速非支配排序方法计算 P_0 中每个 s 的非支配等级 $rk(s)$.
5. 使用拥挤距离分配方法计算 P_0 中每个 s 的拥挤距离 $cd(s)$.
Repeat
6. 应用遗传算子于 P_t 得到子代 O_t.
7. 令 $R_t=P_t\cup O_t$.
8. 使用快速非支配排序方法计算 R_t 中每个 s 的非支配等级 $rk(s)$.
9. 使用拥挤距离分配方法计算 R_t 中每个 s 的拥挤距离 $cd(s)$.
10. 使用拥挤对比算子对 R_t 中的解进行优劣排序得到 R_t'.
11. 将 R_t' 中前 N_p 个体添加到群体 P_{t+1}.
12. 令 $t=t+1$.
Until 终止条件
13. 将终代 P_t 中非支配等级 $rk(s)=1$ 的个体添加到非支配解集 ψ.
14. **Return** ψ

3.2.2 编码方法

编码方法选用二进制编码. 令解 s 的编码为 $\boldsymbol{C}$, 则 $\boldsymbol{C}=(c_1,c_2,\cdots,c_N)$ 为一个 $1\times N$ 的向量. 其中 N 为总特征数, 每个元素 $c_i\in\{0,1\}$ $(i=1,2,\cdots,N)$ 代表第 i 个特征是否被选择, 若为 "1" 则表示被选择, 为 "0" 表示未被选择. 每个编码对应一个解, 也就是一个特征子集.

3.2.3 遗传算子

选择方法. 二进制锦标赛选择为本文所用选择方法. 每次从父代群体 P_t 中选择两个个体, 对比两个个体 (使用拥挤对比算子), 更优者加入子代群体 O_t 中.

交叉算子. 本文选用单点交叉方法. 个体 $\boldsymbol{C_1}=(c_{11},c_{12},\cdots,c_{1N})$、$\boldsymbol{C_2}=(c_{21},c_{22},\cdots,c_{2N})$ 以交叉概率 p_c 进行交叉操作产生两个新个体 $\boldsymbol{C_1}=(c_{11},c_{12},\cdots,c_{1e-1},c_{2e},\cdots,c_{2N})$、$\boldsymbol{C_2}=(c_{21},c_{22},\cdots,c_{2e-1},c_{1e},\cdots,$

c_{1N}), 其中 $e \in \{2,3,\cdots,N\}$ 为随机产生的交叉位.

变异算子. 本文选用单点变异方法. 令个体 $\boldsymbol{C}=(c_1,c_2,\cdots,c_N)$, 则个体中每一位 c_i $(i=1,2,\cdots,N)$ 以变异概率 p_m 进行变异操作, 若原位为 "0", 则变异为 "1", 原位为 "1", 则变异为 "0".

3.3 理想点法

理想点法 [20] 是多目标决策中的常用方法, 本文将理想点法引入特征选择的第二阶段, 从非支配解集中选择最佳调和解作为最终结果. 理想点法分两个步骤完成: 首先, 确定多目标优化的理想点. 接着, 找到与理想点距离最小的解作为最佳调和解. 以下为具体步骤.

1) 归一化处理. 本文多目标优化所确立三个目标函数分别为 f_1 分类错误率, f_2 第 II 类错误率, f_3 特征子集大小. f_1 和 f_2 的取值范围为 $[0,1]$, 而 f_3 的范围要大很多, 因此在应用理想点法之前需对 ψ 中解的三个目标函数值进行归一化处理. 本文采用 z 得分法进行归一化处理, 如式 (4) 所示. 其中 $\overline{f}_i$ 表示第 i 个目标函数的平均值, $\sigma(f_i)$ 表示第 i 目标函数值的标准差.

$$f_i{}^z(s)=\frac{f_i(s)-\overline{f}_i}{\sigma(f_i)},\ i=1,2,3;\ s\in\psi \tag{4}$$

2) 确定理想点. 由于三个目标函数都是望小的, 故确定理想点 F^{z*} 如式 (5) 所示, 其中 f_i^{z*} 为目标函数 $f_i{}^z$ 的最小值, 定义如式 (6).

$$F^{z*}=\{f_1^{z*},f_2^{z*},f_3^{z*}\} \tag{5}$$

$$f_i^{z*}=\min_{s\in\psi}(f_i{}^z(s)),\ i=1,2,3 \tag{6}$$

3) 定义距离. 欧几里德距离简单直观, 本文采用该距离度量解 s 到理想点 F^{z*} 的距离 $d(s)$, 如式 (7) 所示.

$$d(s)=\left[\sum_{i=1}^{3}(f_i^z(s)-f_i^{z*})^2\right]^{1/2},\ s\in\psi \tag{7}$$

4) 选择最佳调和解. 在定义了距离之后, 就可以从 ψ 中选择离理想点最近的解, 作为最佳调和解 s^*, 如式 (8) 所示.

$$s^*=\arg\min_{s\in\psi}(d(s)) \tag{8}$$

综上, 本文算法通过两个步骤进行特征选择. 首先, 通过 NSGA-II 选择非支配解集 ψ. 其次, 通过理想点法从 ψ 中选择出最佳调和解 s^*, 即最优特征子集.

4 算例分析

为了验证算法有效性, 选取 3 个非平衡产品制造数据, 并将算法 CTQ 识别结果与其他 4 个特征选择选法进行对比.

4.1 实验设置

实验数据. 本文选取 3 个非平衡产品生产制造数据进行实验分析, 分别是: 纸产品制造数据 PAPER、胶乳产品制造数据 LATEX 以及尼龙产品制造数据 ADPN [3]. 三个数据集的基本信息如表 3 所示.

表 3 数据集信息

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

对比算法. 本文选取 4 个特征选择算法作为对比算法, 分别是: NSGAII-B、SFS、SBS 以及 PSO. 为了验证本文算法引入第二类错误率的有效性, 实验引入对比算法 NSGAII-B, 该算法仅最小化分类错误率与特征子集大小, 其余均与本文算法相同. SFS 和 SBS 是两个经典的特征选择算法, SFS 通过不断给特征子集中添加有效特征, 达到特征选择的目的, SBS 从原始特征集合中不断剔除无效特征进行特征选择 [8]. PSO 是一类经典随机优化算法, 本文选择基于 PSO 的特征选择算法作为对比算法 [11], 设置终止代数为 500, 粒子数为 100.

实验条件. 本文采用 10 折分层交叉验证方法 [22] 对算法进行验证. 原始数据首先被分为 10 折, 每次取 9 折作为训练集进行 CTQ 识别, 取另外 1 折为测试集验证所识别 CTQ 集的有效性, 最终取 10 次结果的平均作为交叉验证结果. 根据第 2 节分析, 非平衡数据 CTQ 集有效性可以通过分类精度、第 II 类错误率和 CTQ 集大小度量. 因此, 选用这三个指标的交叉验证结果评估 CTQ 识别的效果. 同时, 为了对比各算法的时间复杂度, 实验记录了各算法的运行时间. 本文选取朴素贝叶斯分类器 [23] 作为学习算法, 该算法简洁、高效, 被广泛应用于特征选择领域. SFS、SBS、PSO 算法与朴素贝叶斯分类器使用 Weka[24] 软件实现, 本文算法以及 NSGAII-B 在 Matlab 下实现. 本文算法与 NSGAII-B 使用相同参数设置; 由于 PAPER 数据集包含较少质量特性数, 其染色体编码长度较短, 因此设置群体大小为 100; LATEX 与 ADPN 染色体编码长度较长, 群体大小设置为 200; 交叉概率 p_c 设置为 0.9, 变异概率 p_m 设置为 $1/N$ (N 为总特征数)[19]; 终止代数设置为 500 代. 样本大小的选择: 文献 [25] 提到, 当样本量小于数据维度 (特征数) 时会出现小样本问题. 本文算法初始化时, 个体中每个特征以 0.5 的概率被选择. 从整个群体来看, 有一半 (期望) 个体的特征数小于等于 $N/2$. 那么当样本大小为 $N/2$ 时, 群体中一半个体不遇到小样本问题, 即样本量大于等于特征数. 因此, 为了保证算法性能, 使算法初始化后一半以上的个体避免小样本问题, 本文建议样本大小大于 $N/2$.

4.2 结果与分析

表 4 所示为各特征选择算法所得分类精度, 表中用黑体标出了各行表现最好的算法. 可以看到本文算法在 LATEX 数据集与 ADPN 数据集得到最高的分类精度. 在 PAPER 数据集, SBS 分类精度最高, NSGAII-B 与本文算法表现略差于 SBS. 本文算法在各数据集的平均分类精度达到 84.98%, NSGAII-B 表现其次, 达到 83.64%, SBS、PSO 和 SFS 表现接近, 分别是 80.45%、80.32%、79.33%. 综合来看, 本文算法所得 CTQ 集能够得到不错的分类精度.

表 5 所示为各特征选择算法所得第 II 类错误率, 各行表现最好的算法用黑体标出. 可以看到本文算法在 LATEX 数据集与 ADPN 数据集得到了最低的第 II 类错误率, 分别为 24.11%、10.00%. 在 PAPER 数据集, SBS 得到了最低的第 II 类错误率, 本文算法略高于 SBS, 其次为 PSO 、NSGAII-B 和 SFS. 根据平均结果, 本文算法的第 II 类错误率最低, PSO 算法次之, 最后为 SBS、NSGAII-B 和 SFS. 综合来看, 相对于其他算法, 本文算法显著降低了第 II 类错误率. 说明通过在特征选择阶段引入第 II 类错误率作为优化目标, 算法能够有效处理数据不平衡性.

表 6 所示为各算法所得 CTQ 集大小, 各行表现最好的算法用黑体标出. 可以看到本文算法在 PAPER 和 ADPN 数据集得到最小 CTQ 集, NSGAII-B 在 LATEX 数据集得到最小 CTQ 集. 根据平均结果, 本文算法所得 CTQ 集大小为 3.5, NSGAII-B 为 3.3, SFS 和 PSO 分别为 5.8 和 14.6, SBS 表现最差, 为 47.9. 本文算法和 NSGAII-B 的降维能力要好于其余 3 个对比算法, 说明多目标进化算法 NSGA-II 在特征选择时有更好的降维能力; 通过将特征子集大小单独作为一个优化目标, 特征维度能够显著降低.

表 7 所示为各算法运行时间, 各行运行时间最短的算法用黑体标出. 可以看到 SFS 在各数据集的运行时间远小于其他算法, 其次为 SBS. 本文算法与 NSGAII-B 在各数据的运行时间接近. PSO 在 LATEX 与 ADPN 上的运行时间略高于本文算法, 在 PAPER 数据集运行时间略低与本文算法. 总体来看, 本文算法、NSGAII-B、PSO 与 SBS 的运行时间处于同一数量级, 而 SFS 运行时间要远小于其他算法.

表 4　各算法分类精度 (%) 对比

数据集	本文算法	NSGAII-B	PSO	SFS	SBS
PAPER	87.23	88.02	83.08	82.06	**89.30**
LATEX	**83.25**	81.47	80.57	78.62	76.35
ADPN	**84.46**	81.42	77.32	77.32	75.71
平均	**84.98**	83.64	80.32	79.33	80.45

表 5　各算法第 II 类错误率 (%) 对比

数据集	本文算法	NSGAII-B	PSO	SFS	SBS
PAPER	21.67	34.17	26.67	41.67	**19.17**
LATEX	**24.11**	30.54	28.39	50.18	38.75
ADPN	**10.00**	30.00	25.00	25.00	35.00
平均	**18.59**	31.57	26.69	38.95	30.97

表 6 各算法所得 CTQ 集大小对比

数据集	本文算法	NSGAII-B	PSO	SFS	SBS
PAPER	**2.6**	4.2	5.6	4.1	21.6
LATEX	5.6	**3.3**	27.9	8.0	96.3
ADPN	**2.2**	2.5	10.3	5.3	25.8
平均	3.5	**3.3**	14.6	5.8	47.9

表 7 各算法运行时间 (s) 对比

数据集	本文算法	NSGAII-B	PSO	SFS	SBS
PAPER	39851	36392	14490	**516**	16161
LATEX	131444	103579	202064	**4590**	61622
ADPN	17366	17068	26413	**500**	15134
平均	62887	52346	80989	**1869**	30972

综合以上结果, 可以看到本文算法能够在各数据集上获得比较高的分类精度, 同时能够有效降低第 II 类错误率, 并能有效进行特征降维得到较小 CTQ 集. 从时间复杂度来看, 本文算法高于 SFS 算法, 与其他对比算法接近. 总体来说, 本文算法能够针对不平衡产品制造数据, 有效过滤无关、冗余质量特性, 识别产品 CTQ.

5 结论

为了解决非平衡产品制造数据 CTQ 识别问题, 本文提出基于 NSGA-II 的非平衡数据特征选择算法. 首先, 针对传统特征选择算法处理非平衡数据的不足, 本文引入第 II 类错误率作为非平衡数据特征子集重要性的度量, 最终确立优化目标为最小化分类错误率、最小化第 II 类错误率以及最小化特征子集大小. 该处理能够保证所选特征子集对少类样本的分类能力. 接着, 本文采用多目标进化算法 NSGA-II 优化以上三个度量标准, 得到一个非支配解集. 最后, 为了解决非支配解集包含过多解的问题, 本文引入理想点法, 从所得非支配解集中选择最佳调和解, 作为最终的特征子集. 算例分析阶段, 所提算法被应用于 3 组非平衡产品制造数据的 CTQ 识别. 实验结果表明, 本文算法在保证较高分类精度的同时, 有效降低了第 II 类错误率和 CTQ 集大小; 说明算法能够有效针对数据非平衡性, 过滤无关、冗余质量特性, 进行 CTQ 识别. 降低算法的时间复杂度, 针对小样本制造数据提出 CTQ 识别算法, 是今后需要做的工作.

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 Technical Conference, 1996.

[2] 闫伟, 何桢, 李岸达. 基于 CEM-IG 算法的复杂产品关键质量特性识别 [J]. 系统工程理论与实践, 2014, 34(5): 1230–1236.
Yan W, He Z, Li A D. Identification of critical-to-quality characteristics for complex products using CEM-IG algorithm[J]. Systems Engineering — Theory & Practice, 2014, 34(5): 1230–1236.

[3] 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.

[4] Jeong B, Cho H. Feature selection techniques and comparative studies for large-scale manufacturing processes[J]. The International Journal of Advanced Manufacturing Technology, 2006, 28(9–10): 1006–1011.

[5] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy[J]. The Journal of Machine Learning Research, 2004, 5: 1205–1224.

[6] Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF[J]. Machine Learning, 2003, 53(1–2): 23–69.

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

[8] Gunal S, Gerek O N, Ece D G, et al. The search for optimal feature set in power quality event classification[J]. Expert Systems with Applications, 2009, 36(7): 10266–10273.

[9] Pacheco J, Casado S, Angel-Bello F, et al. Bi-objective feature selection for discriminant analysis in two-class classification[J]. Knowledge-Based Systems, 2013, 44: 57–64.

[10] 姚旭, 王晓丹, 张玉玺, 等. 基于粒子群优化算法的最大相关最小冗余混合式特征选择方法 [J]. 控制与决策, 2013, 28(3): 413–417.
Yao X, Wang X D, Zhang Y X, et al. A maximum relevance minimum redundancy hybrid feature selection

algorithm based on particle swarm optimization[J]. Control and Decision, 2013, 28(3): 413–417.
[11] Moraglio A, Di Chio C, Togelius J, et al. Geometric particle swarm optimization[J]. Journal of Artificial Evolution and Applications, 2008, 2008: 1–14.
[12] Huang C L, Dun J F. A distributed PSO-SVM hybrid system with feature selection and parameter optimization[J]. Applied Soft Computing, 2008, 8(4): 1381–1391.
[13] Huang C L, Wang C J. A GA-based feature selection and parameters optimization for support vector machines[J]. Expert Systems with Applications, 2006, 31(2): 231–240.
[14] 邬开俊, 鲁怀伟. 采用并行协同进化遗传算法的文本特征选择 [J]. 系统工程理论与实践, 2012, 32(10): 2215–2220.
Wu K J, Lu H W. PCEGA used to solve text feature selection[J]. Systems Engineering — Theory & Practice, 2012, 32(10): 2215–2220.
[15] Vignolo L D, Milone D H, Scharcanski J. Feature selection for face recognition based on multi-objective evolutionary wrappers[J]. Expert Systems with Applications, 2013, 40(13): 5077–5084.
[16] Borgelt C, Gil M Á, Sousa J, et al. Towards advanced data analysis by combining soft computing and statistics[M]. New York: Springer, 2013: 359–375.
[17] Xue B, Zhang M, Browne W N. Multi-objective particle swarm optimisation (PSO) for feature selection[C]// Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation, ACM, 2012: 81–88.
[18] Su C T, Chen L S, Chiang T L. A neural network based information granulation approach to shorten the cellular phone test process[J]. Computers in Industry, 2006, 57(5): 412–423.
[19] Deb K, Pratap A, Agarwal S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II[J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182–197.
[20] Freimer M, Yu P L. Some new results on compromise solutions for group decision problems[J]. Management Science, 1976, 22(6): 688–693.
[21] Srinivas N, Deb K. Muiltiobjective optimization using nondominated sorting in genetic algorithms[J]. Evolutionary Computation, 1994, 2(3): 221–248.
[22] Han J, Kamber M, Pei J. Data mining: Concepts and techniques[M]. 3rd ed. Waltham, MA: Morgan Kaufmann, 2012: 364–377.
[23] 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.
[24] 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.
[25] Chen L F, Liao H Y M, Ko M T, et al. A new LDA-based face recognition system which can solve the small sample size problem[J]. Pattern Recognition, 2000, 33(10): 1713–1726.