第 37 卷 第 9 期
2015 年 9 月

系统工程与电子技术
Systems Engineering and Electronics

Vol. 37 No. 9
September 2015

文章编号:1001-506X(2015)09-2073-07

网址:www.sys-ele.com

基于 GSA 的复杂产品关键质量特性识别

李岸达,何 桢,何曙光
(天津大学管理与经济学部,天津 300072)

摘 要:为了识别复杂产品关键质量特性(critical-to-quality characteristics,CTQs),提出基于遗传模拟退火算法(genetic simulated annealing algorithm, GSA)的特征选择算法。所提算法将遗传算法(genetic algorithm, GA)与模拟退火算法(simulated annealing algorithm, SA)结合,兼有不错局部搜索与全局搜索能力。提出一种综合适应度函数应用于所提算法,以同时优化 CTQ 集分类性能和所选质量特性数。算例结果表明,所提算法能有效过滤无关、冗余质量特性,识别关键质量特性;与 Memetic 算法和信息增益(information gain, IG)算法相比,所提算法在识别更少关键质量特性的同时,得到更高预测精度。

关键词:关键质量特性;遗传算法;模拟退火算法;复杂产品;特征选择

中图分类号:F 406.3 **文献标志码**:A **DOI**:10.3969/j.issn.1001-506X.2015.09.18

Critical-to-quality characteristics identification for complex products using GSA

LI An-da, HE Zhen, HE Shu-guang
(*College of Management and Economics, Tianjin University, Tianjin 300072, China*)

Abstract: To identify critical-to-quality characteristics (CTQs) for complex products, a genetic simulated annealing algorithm(GSA) based feature selection algorithm is proposed. As the proposed algorithm combines the genetic algorithm (GA) and simulated annealing algorithm (SA), it has both good local search ability and good global search ability. Additionally, the proposed algorithm adopts an aggregated fitness function, which can optimize the classification performance on CTQ set and the number of selected quality characteristics simultaneously. Experimental results illustrate that the proposed algorithm can efficiently eliminate irrelevant and redundant quality characteristics and identify CTQs, as it can identify fewer CTQs with even higher predictive accuracy compared with the Memetic algorithm and the information gain (IG) algorithm.

Keywords: critical-to-quality characteristics (CTQs); genetic algorithm (GA); simulated annealing algorithm (SA); complex products; feature selection

0 引 言

具有“客户需求复杂、产品组成复杂、产品技术复杂、制造流程复杂、试验维护复杂、项目管理复杂、工作环境复杂”等特征的一类产品被称之为复杂产品[1]。由于复杂产品零部件众多,结构复杂,质量特性相互影响关系复杂,因此对该类产品进行质量控制、质量监控的难度相对较高。在实际生产中,由于不能判断复杂产品制造过程中哪些质量特性是关键质量特性(critical-to-quality characteristics, CTQs),只能将所有零件公差范围收缩,并对每一个生产过程进行严格监控,从而导致加工成本增加,生产周期变长。对于复杂产品,如何从众多质量特性中有效过滤无关、冗余质量特性,识别影响产品质量的 CTQ 是一个亟待解决的问题。

传统 CTQ 识别方法包括关键特性展开(key characteristics flowdown, KCF)[2]和质量功能展开(quality function deployment, QFD)[3]。KCF 对产品进行逐层分解,将产品从上到下分解为产品特性、部件特性、零件特性、工艺特性等,之后应用定性、定量的分析方法,从中识别 CTQ[2,4]。但是复杂产品包含大量零部件级质量特性,质量特性间影响关系复杂,难以用传统定性、定量方法确定各质量特性间的影响关系并识别 CTQ。QFD 把顾客对产品的要求进行多层分析,最终转换为产品的设计要求、生产要求、工艺要求等,从而建立产品的生产策略,该方法的重要特点是体现了客户需求导向[3]。但是顾客只关注对其使用有直接影响的因素,很多影响产品(尤其是复杂产品)质量潜在因素是顾客所不能关注到的。此外,对于复杂产品,在高纬度带来

收稿日期:2014-07-08; 修回日期:2014-12-25; 网络优先出版日期:2015-01-20。
网络优先出版地址:http://www.cnki.net/kcms/detail/11.2422.TN.20150120.1050.005.html
基金项目:国家自然科学基金(71102140);国家杰出青年科学基金(71225006)资助课题

的复杂性影响下，QFD 的质量矩阵不易确定，最终影响 CTQ 识别效果[5]。

特征选择是机器学习领域一类能够有效处理高维数据集的降维方法[6-7]。该类方法相继被引入复杂产品 CTQ 识别领域[5,8]。通常，特征选择算法可以分为两类：Filter 算法和 Wrapper 算法。

在 Filter 算法中，特征选择是在应用学习算法分类之前的一个预处理步骤，该类算法通过一定的评估策略过滤掉一些特征，接着应用学习算法得到算法分类精度。文献[5]应用一种经典的过滤算法信息增益（information gain，IG）进行关键质量特性识别，该方法以信息增益为度量标准，计算各质量特性（特征）与产品质量（类标签）之间的相关程度，最终得到各质量特性的权重，然后过滤掉权重较小的质量特性，得到 CTQ 集。但是，由于 IG 算法单独评价每个质量特性的重要性，有些潜在重要的质量特性不能发现；另外 IG 算法没有考虑质量特性间的冗余性，不能有效过滤冗余质量特性。

Wrapper 算法将学习算法包含在特征选择过程中，学习算法的分类性能是评价特征子集好坏的重要指标。Wrapper 算法在进行特征选择时将每个特征集合作为一个整体对其好坏进行评价的，能够发现潜在关键特征和处理特征间的冗余性[9]。Wrapper 算法可以看作是一个优化过程，其目标通常是选择一个尽可能小的特征子集并使它的分类性能尽可能大，所以 Wrapper 算法通常包含一个搜索策略。Wrapper 算法常用的搜索策略包括各类顺序寻优算法[10]以及遗传算法[11-12]（genetic algorithm，GA）。相对于顺序寻优算法，全局最优算法 GA 能够脱离局部最优，并在更广的空间内搜索特征子集[12]。文献[13]将特征选择作为一个单目标优化问题，以分类性能作为 GA 算法的适应度函数。然而，仅优化分类性能容易造成对训练集的过拟合，最终会选择过多的特征，不利于过滤无关、冗余特征。因此，将最小化特征数也作为优化目标以有效处理以上问题是必要的[14]。

GA 算法能够有效地全局寻优，所以将其应用于复杂产品 CTQ 识别能够有效解决高维度问题。但是，GA 算法局部搜索能力较差，单独使用 GA 算法不能得到理想结果。Memetic 算法可以看作是一种混合 GA 算法，它将局部搜索与 GA 算法相结合，弥补了 GA 算法局部搜索能力差的缺点，通常比 GA 算法有更好效果[15]。由于模拟退火（simulated annealing algorithm，SA）算法有较强的局部搜索能力，部分学者将 SA 与 GA 结合，建立遗传模拟退火算法（genetic simulated annealing algorithm，GSA），该算法同样能够弥补 GA 算法的缺点，快速收敛得到满意解[16-17]。

基于以上分析，本文将 GSA 应用于特征选择，构建基于 GSA 的特征选择算法，并将其应用于复杂产品 CTQ 识别。所提算法应用了一种综合适应度函数以同时优化 CTQ 集的分类性能和所选质量特性数。算例结果表明，相对于 IG 算法与 Memetic 算法，所提算法能够在得到更高预测精度的同时，识别更少 CTQ。说明所提算法能够有效过滤无关、冗余质量特性。

1 基于 GSA 的 CTQ 识别框架

假设在产品质量特性数据集 Ω 中，包含 M 个样本，每个样本可以表示为 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{iK},y_i)$，$i=1,\cdots,M$；产品样本含有 K 个质量特性，分别用 Q_i 表示，$i=1,\cdots,K$；x_{ij} 表示第 i 个产品样本的第 j 个质量特性的测量值；每个产品样本的类标签是 y_i。类标签表示每个产品样本的质量好坏。例如，若产品分为“合格”、“不合格”两类，则可以分别赋予类标签为“-1”和“1”。产品样本 $\boldsymbol{X}_i$ 最终属于哪个类别受到其质量特性取值 $x_{i1},x_{i2},\cdots,x_{iK}$ 的影响，但是不是每个质量特性都是影响产品质量的关键影响因素，需要从这些质量特性中识别 CTQ。本文应用特征选择算法识别 CTQ，由于特征选择算法能够选择出影响样本分类最显著的特征，故可以将特征选择算法用于识别影响产品质量（分类）的关键质量特性（特征）[8]。

由于 GSA 算法兼有不错全局与局部寻优的能力，本文提出基于 GSA 算法的特征选择算法，并将该算法用于复杂 CTQ 识别。基于 GSA 算法的 CTQ 识别框架如图 1 所示。识别框架可以分为 3 个阶段。第一阶段：数据集划分。将数据集划分为产品训练数据集与产品测试数据集。第二阶段：利用 GSA 算法识别 CTQ。将产品训练数据集输入基于 GSA 的特征选择算法，识别得到 CTQ 集。第三阶段：CTQ 识别结果评估。使用测试集得到 CTQ 集的预测精度，并评估 CTQ 识别结果。CTQ 集预测精度与 CTQ 集包含质量特性数是 CTQ 识别效果的两个重要度量标准，预测精度越高，质量特性数越少，说明算法能够越有效过滤无关、冗余质量特性，从而有效识别 CTQ。识别框架具体步骤如下：

第一阶段：

步骤 1 从生产制造过程中收集产品质量特性数据集。

步骤 2 将原始数据集划分为两部分：产品质量特性训练数据集与产品质量特性测试数据集。

第二阶段：

步骤 3 将训练集输入基于 GSA 的特征选择算法，对 GSA 算法进行初始化参数设置，令 $i=0$。

步骤 4 产生初始群体 G_0，并评价 G_0 各个体的适应度。

步骤 5 对群体 G_i 进行遗传操作（包括选择、交叉、变异），得到过渡群体 G'_{i+1}，并评价 G'_{i+1} 各个体的适应度。

步骤 6 对 G'_{i+1} 的个体进行模拟退火操作，根据 Metropolis 准则判断是否接受从 G_i 中父代个体到 G'_{i+1} 中子代个体的转移，得到群体 G_{i+1}，并评价 G_{i+1} 个体的适应度。

步骤 7 判断是否满足 GSA 算法的终止条件，若满足终止条件则进入步骤 8；否则，令 $i=i+1$，并返回步骤 5。

步骤 8 从群体中 G_{i+1} 选择具有最高适应度的个体作

为最终结果，得到 CTQ 集。

第三阶段：

步骤 9　使用训练集训练学习算法，并使用测试集测试训练后学习算法，得到 CTQ 集对应的预测精度。

步骤 10　根据预测精度及 CTQ 集质量特性数评估 CTQ 集。

2　基于 GSA 的特征选择算法

第 1 节提出基于 GSA 的 CTQ 识别框架，本节提出“基于 GSA 的特征选择算法”（见图 1 虚框）。本节余下部分包括：编码方式、初始群体的产生、适应度函数、选择操作、交叉操作、变异操作、模拟退火操作、终止条件。

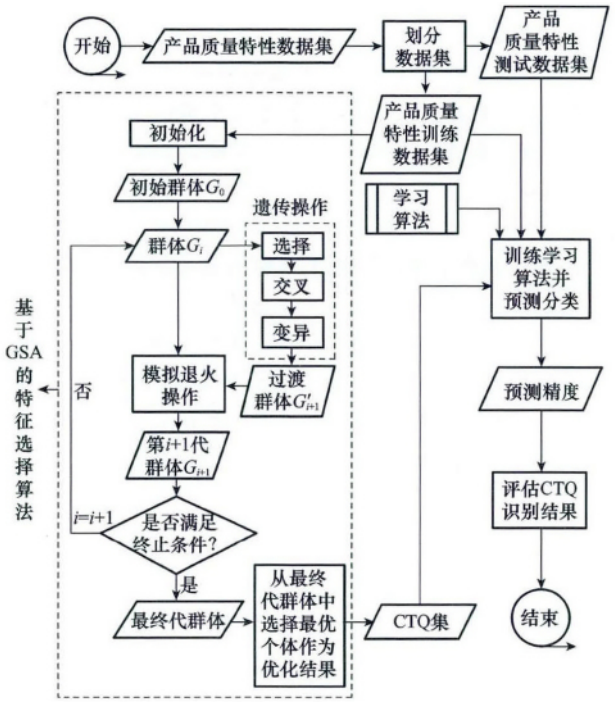

图 1　GSA 算法 CTQ 识别框架

2.1　编码方式

编码采用二进制编码方式，若产品质量特性数据集 Ω 包含 K 个质量特性，则个体编码 $\boldsymbol{B}=(b_1,b_2,\cdots,b_K)$，其中 $b_j\in\{0,1\}$，$j=1,2,\cdots,K$，编码长度 $N_B=K$。编码中每一个 b_j 表示对应的第 j 个质量特性是否被识别为 CTQ。若 $b_j=1$，则表示第 j 个质量特性被包含在 CTQ 集中；若 $b_j=0$，则表明第 j 个质量特性不在 CTQ 集中。

若一个数据集中包含 10 个质量特性，第 1、3、5、9 个质量特性被包含在 CTQ 集中，则个体编码为 $\boldsymbol{B}=(1,0,1,0,1,0,0,0,1,0)$。解码过程与编码过程相反，若个体编码 $\boldsymbol{B}=(0,0,1,1,0,0,1,0,1,0)$，则表明第 3、4、7、9 个质量特性被包含在 CTQ 集中。

2.2　初始群体的产生

随机产生初始群体，产生 $1\times K$ 的随机向量，向量中每一个元素 $b_j(j=1,2,\cdots,K)$ 随机取值为 0 或 1，令其取 1 的概率为 P_1，则初始化方法如式(1)所示，其中 $Rand(0,1)$ 表示以均匀分布在 0 到 1 之间产生的随机数。由于复杂产品数据集维度较高，为了使算法更快收敛到较小的质量特性集，在初始化时选择一个较小的 P_1，本文选取 $P_1=0.3$。群体规模 N_p 与问题的复杂程度有关，通常在 50 到 500 之间[11]，本文选取 $N_p=100$。

$$b_j=\begin{cases}0, & Rand(0,1)\geqslant P_1\\ 1, & Rand(0,1)<P_1\end{cases};\ j=1,2,\cdots,K \tag{1}$$

2.3　适应度函数

本节提出一个改进综合适应度函数，以同时优化 CTQ 集（特征子集）分类性能和所选质量特性（特征）数。文献[13]在特征选择时仅将最大化特征子集的分类性能作为优化目标。但是由于缺少对所选特征数的优化，算法会造成对训练集的过拟合，不能有效过滤冗余、无关特征等问题[14]。为了解决以上问题，本文将所选质量特性数引入适应度函数，构造一个综合适应度函数，以同时优化 CTQ 集分类性能和所选质量特性数。应用层面，识别少量 CTQ 对于节约成本、有效提高产品质量是很有意义的。这样可以将有限时间与金钱用于少量影响产品质量的关键特性，从而能够显著改善产品质量并节约成本。所以，从理论与实际两方面来看，将最小化质量特性数作为优化目标很有必要。

根据以上分析，可知在进行 CTQ 识别时有两个优化目标：①最大化 CTQ 集分类性能；②最小化所选质量特性数。通常，分类性能的估计是通过训练集的内部 5 折交叉验证精度得到的[9]。所选质量特性个数可以通过计算个体编码“1”的个数得到。

优化第一个目标，可以建立如式(2)所示的目标函数。

$$J_1(\boldsymbol{B}_i)=R(QS(\boldsymbol{B}_i)) \tag{2}$$

式中，$QS(\boldsymbol{B}_i)$ 是个体 $\boldsymbol{B}_i$ 对应的质量特性集；$R(QS(\boldsymbol{B}_i))$ 表示 $QS(\boldsymbol{B}_i)$ 对应的 5 折交叉验证精度。

优化第二个目标，可以建立如式(3)所示的目标函数，该式表示了被过滤掉的质量特性数所占总特性数的比例。

$$J_2(\boldsymbol{B}_i)=\left(1-\frac{\#QS(\boldsymbol{B}_i)}{\#FS}\right) \tag{3}$$

式中，$\#QS(\boldsymbol{B}_i)$ 表示质量特性子集 $QS(\boldsymbol{B}_i)$ 包含质量特性的个数；$\#FS$ 代表原质量特性数。

通过最大化式(3)可以最小化所选质量特性数。在式(2)与式(3)的基础上，可以将两个目标综合，建立如式(4)所示的综合目标函数。通过优化式(4)，能够同时优化两个目标。

$$J_3(\boldsymbol{B}_i)=\beta\cdot R(QS(\boldsymbol{B}_i))+(1-\beta)\cdot\left(1-\frac{\#QS(\boldsymbol{B}_i)}{\#FS}\right),\quad 0\leqslant\beta\leqslant 1 \tag{4}$$

式中，β 是一个调节参数，主要目的是调整分类性能与所选质量特性数的相对重要程度，β 取值在 0 到 1 之间，β 取值越大交叉验证精度越重要，当 $\beta=0$ 时，只考虑质量特性数，当 $\beta=1$ 时，只考虑交叉验证精度。

J_3 可以同时优化分类精度和质量特性数，但是直接将 J_3 作为适应度函数也会有以下问题。由于 J_3 的取值范围

是在 0 到 1 之间，不同的个体 B_i 对应的 J_3 目标函数值差异可能是很小的，GSA 算法在进行选择操作的时候就不能有效的将目标函数值高的个体选择出来，从而影响 GSA 算法的收敛速度和算法寻优结果。所以通过一种变换适当放大不同个体适应度的差异是必要的。本文引入指数函数，将 J_3 变换得到 J_4，如式(5)所示。

$$J_4(\boldsymbol{B}_i) = \mathrm{e}^{\alpha * J_3(\boldsymbol{B}_i)}, \alpha > 0 \tag{5}$$

式中，α 是放大差异参数。令 $\boldsymbol{B}_i$、$\boldsymbol{B}_j$ 为两不同个体，且 $J_3(\boldsymbol{B}_i) > J_3(\boldsymbol{B}_j)$，则有式(6)推导，可以看到 $J_4(\boldsymbol{B}_i)$ 与 $J_4(\boldsymbol{B}_j)$ 的比例是与 α 密切相关的，α 越大这个比例就越大，所以 α 能够有效放大不同个体目标函数值的差异。通过以上分析，可知目标函数 J_4 既能同时优化分类性能和质量特性数，又能放大不同个体好坏的差异。

$$J_4(\boldsymbol{B}_i)/J_4(\boldsymbol{B}_j) = \exp(\alpha \cdot (J_3(\boldsymbol{B}_i) - J_3(\boldsymbol{B}_j))), \alpha > 0 \tag{6}$$

根据以上分析，本文将 J_4 作为 GSA 的适应度函数，该适应度函数是一个综合适应度函数，可以同时最大化分类性能和最小化所选质量特性数，如式(7)所示。

$$fitness(\boldsymbol{B}_i) = \exp(\alpha \cdot (\beta \cdot R(QS(\boldsymbol{B}_i)) + (1-\beta) \cdot (1 - \frac{\#QS(\boldsymbol{B}_i)}{\#FS}))),$$
$$\alpha > 0;\ 0 \leqslant \beta \leqslant 1 \tag{7}$$

式中，$R(QS(\boldsymbol{B}_i))$ 为质量特性集 $QS(\boldsymbol{B}_i)$ 对应的训练集 5 折交叉验证精度；$\#QS(\boldsymbol{B}_i)$ 表示 $QS(\boldsymbol{B}_i)$ 包含质量特性的个数；$\#FS$ 代表原始数据集中包含的质量特性的个数；α 是放大参数，能够放大适应度的差异；β 是一个调节参数，可以调整 CTQ 集分类性能与所选质量特性数的相对重要程度，β 取值在 0 到 1 之间，β 取值越大分类性能权重越高，反之亦然。本文取 $\alpha=4$，$\beta=0.7$。

2.4 选择操作

采用轮盘赌方式进行选择操作，将群体中的个体按照其适应度从大到小排序，然后进行选择操作。个体被选择的概率和其适应度有关，适应度越大则被选择的概率就越大，反之亦然。采用轮盘赌方式从父代 G_i 中选择 N_p 个个体，并对选择个体配对得到 $N_p/2$ 对个体，然后进入交叉操作。

2.5 交叉操作

采用一点交叉方式进行交叉操作。个体以交叉概率 p_c 进行交叉操作。对每对被选择的个体，随机产生一个交叉位(不能是最后一位)，两个个体交叉位之后的编码串互换。例如，对于个体 $\boldsymbol{B}_1=(b_{11},b_{12},\cdots,b_{1K})$ 和 $\boldsymbol{B}_2=(b_{21},b_{22},\cdots,b_{2K})$，随机选择交叉位 $x\in\{1,2,\cdots,K-1\}$，交换之后产生两个新的个体 $\boldsymbol{B}_1'=(b_{11},b_{12},\cdots,b_{1x},b_{2(x+1)},\cdots,b_{2K})$，$\boldsymbol{B}_2'=(b_{21},b_{22},\cdots,b_{2x},b_{1(x+1)},\cdots,b_{1K})$。本文选取 $p_c=0.9$。

2.6 变异操作

本文采用一种子集导向变异算子(subset size oriented mutation，SSOM)，在应用于特征选择算法时，该变异算子有更好效果[18]。SSOM 能够保证“0”位与“1”位变异的期望个数相等，这能够保证变异算子整体上不改变变异之前“0”位与“1”位的比例，也就是变异过程中从特征集合剔除的特征数与包含到特征集合内的特征数的期望是相等的。SSOM 对“0”位的变异与“1”位的变异是分别进行的，令 p_{m1} 为个体中“1”位变异的概率，则个体中“0”位变异的概率 p_{m0} 由下式得到：

$$p_{m0} = \frac{N_1}{N_0} p_{m1} \tag{8}$$

式中，N_1 表示变异前个体中“1”位的个数；N_0 表示变异前个体中“1”位的个数；本文取 $p_{m1}=0.01$。

2.7 模拟退火操作

模拟退火算法优化目标函数是望小的，但是遗传算法的适应度函数是望大的，所以在进行模拟退火操作之前需要对目标函数进行变换。式(9)为变换后的模拟退火算法目标函数。

$$f(\boldsymbol{B}) = v - fitness(\boldsymbol{B}) \tag{9}$$

式中，$\boldsymbol{B}$ 代表群体中的个体；f 为模拟退火算法优化目标函数。通过式(9)的变换，将最大值优化问题改为最小值优化问题，v 为足够大的数保证变换后的目标函数非负。

令 $\boldsymbol{d}_1$，$\boldsymbol{d}_2$ 为父代 G_i 的一对个体，该对个体经过交叉、变异之后，得到过渡代 G'_{i+1} 的一对个体 $\boldsymbol{s}_1$，$\boldsymbol{s}_2$。那么以概率 P 接受 $\boldsymbol{s}$ 为群体 G_{i+1} 中的个体，如式(10)所示。

$$P = \begin{cases} 1, & f(\boldsymbol{s}) < f(\boldsymbol{d}) \\ \exp(-\dfrac{f(\boldsymbol{s}) - f(\boldsymbol{d})}{T}), & f(\boldsymbol{s}) \geqslant f(\boldsymbol{d}) \end{cases} \tag{10}$$

将式(9)代入式(10)得到概率 P 的计算公式如下：

$$P = \begin{cases} 1, & fitness(\boldsymbol{s}) > fitness(\boldsymbol{d}) \\ \exp(\dfrac{fitness(\boldsymbol{s}) - fitness(\boldsymbol{d})}{T}), & fitness(\boldsymbol{s}) \leqslant fitness(\boldsymbol{d}) \end{cases} \tag{11}$$

式(10)、式(11)中 T 是模拟退火算法中的温度，算法运行中温度逐渐降低。令第 i 代温度为 T_i，则 $T_{i+1}=\eta \cdot T_i$，$0<\eta<1$。本文取初始温度 $T_0=2\,000$，系数 $\eta=0.95$。

进行模拟退火操作之后，得到下一代群体 G_{i+1}，为了保证算法不丢失所得到的最优解，采用精英保留策略，将 G_i 代中的最优个体保留并加入到群体 G_{i+1} 中。

2.8 终止条件

群体进化 N_t 代之后，进化操作停止，选取最后一代有最高适应度值的个体作为最终解。本文选取 $N_t=500$。

3 算法应用实例

为了验证算法有效性，本节选取 3 个数据集进行实验，分别是 AIRMANU、SPIRA 以及 LATEX。数据集 AIR-

MANU 是飞机陀螺仪的质量特性数据集，其质量特性主要包含陀螺仪的物理参数、光学参数、电参数等，该数据收集自国内某航空研究所。飞机陀螺仪结构复杂，质量特性众多，在实际生产中进行质量控制难度较大，导致产品合格率较低，有大量的返修情况。因此，识别陀螺仪 CTQ，并对所识别 CTQ 进行重点控制是一项非常重要的任务。数据集 SPIRA 收集自抗生素生产过程，产品质量特性包括温度水平和耗氧峰值等；数据集 LATEX 收集自胶乳生产过程，产品质量特性包括反应浓度、温度水平等[19]。同样，从大量质量特性中识别 CTQ 并进行重点控制，对于提高抗生素以及胶乳的质量有重要意义。AIRMANU、SPIRA 以及 LATEX 数据集信息如表 1 所示。

表 1　数据集信息

数据集	样本数	合格品/ 不合格品	质量特性数
AIRMANU	87	45/42	117
SPIRA	145	95/50	96
LATEX	262	184/78	117

实施条件。在进行 CTQ 识别之前，需要将数据集划分为训练集和测试集。训练集用来进行 CTQ 识别，测试集用来验证 CTQ 识别的有效性。要使算法有效识别 CTQ，应保证训练集中有足够的训练数据，因此本文将数据集按照 3∶1的比例划分为训练集和测试集。朴素贝叶斯分类器被广泛用于特征选择领域[20]，因此本文选取朴素贝叶斯分类器作为学习算法。为了验证算法有效性，本文选取两个对比算法，第一个是经典 Filter 算法——IG 算法[5]；第二个是一种 Wrapper 算法——Memetic 算法，一种混合遗传算法[15]。本文算法与 Memetic 算法的参数设置如表 2 所示。此外，IG 算法使用 Weka[21]实现；本文算法与 Memetic 算法在 Matlab 环境下编程实现，所用朴素贝叶斯分类器从 Weka 中调用。

表 2　本文算法与 Memetic 算法参数设置

参数	算法	
	本文算法	Memetic 算法
群体规模 N_p	100	100
交叉概率 p_c	0.9	0.9
变异概率(p_{m1}/p_m)	0.01 (p_{m1})	0.01 (p_m)
终止代数 N_t	500	500
初始温度 T_0	2 000	—
系数 η	0.95	—

本文应用两个标准度量 CTQ 的识别效果[7]：①CTQ 集的测试集预测精度。预测精度越高，CTQ 集的相关性越好，说明影响产品质量的关键特性被包含在 CTQ 集中。②CTQ 集所选质量特性数。质量特性数越少，则表明 CTQ 识别算法越多地有效了过滤无关、冗余的质量特性。综合来看，CTQ 集的预测精度越高，包含质量特性越少，CTQ 识别越有效。

图 2 所示为本文算法与 Memetic 算法的收敛性能图，各子图分别表示了 3 个数据集的训练集 5 折交叉验证精度(为简要表述之后用交叉验证精度代替)收敛曲线与所选质量特性数收敛曲线。由交叉验证精度收敛曲线(见图 2(a)、图 2(c)和图 2(e))可以看到，两个算法都能有效收敛。相比本文算法，Memetic 算法在 AIRMANU 上得到了更高交叉验证精度。在 SPIRA 与 LATEX 上，两个算法最终得到了相等交叉验证精度。由所选质量特性数收敛曲线(见图 2(b)、图 2(d)、图 2(f))可以看到，本文算法能够快速降低所选质量特性数，并达到收敛状态；Memetic 算法同样能够达到收敛，但是质量特性数降低效果要差与本文算法。造成以上结果，是因为本文算法在进行特征选择时，是综合优化交叉验证精度与所选质量特性数两个指标的，所以本文算法在提高交叉验证精度的同时，显著降低了所选质量特性数。Memetic 单纯优化交叉验证精度，能够得到较高交叉验证精度，但是由于单纯优化交叉验证精度，造成了过多质量特性被选择。另外，从收敛速度来说，本文算法在 3 个数据集上都能在 150 代之前达到交叉验证精度与所选质量特性数的收敛，有不错的收敛速度。综合来说，本文算法能够快速有效收敛。

表 3 所示为各识别算法的 CTQ 识别结果。由表 3 可知，本文算法在 3 个数据集都有不错预测精度，平均预测精度达到 85.77%。IG 算法与 Memetic 算法在 3 个数据集上的预测精度都不高于本文算法，平均预测精度分别为 77.44%和 79.04%。同时，在 3 个数据集上，本文算法能够选择更少关键质量特性，平均选择质量特性为 11.33 个。IG 与 Memetic 算法平均选择质量特性分别为 29.00 和29.67个。综合来看，本文算法能够在得到更高预测精度的同时选择更少的质量特性。得到这样的结果，是因为本文算法将 CTQ 集所选质量特性数作为一个优化目标。一方面，通过对质量特性的数量的控制，算法尽可能少的选择质量特性，将最重要的质量特性保留下来，从而达到过滤无关、冗余质量特性的目的。另一方面，将最小化 CTQ 集所选质量特性数包含到适应度函数，能够同时优化 CTQ 集的分类性能和所选质量特性数，使算法避免了一味追求训练集交叉验证精度最大化而造成对训练集的过拟合。Memetic 算法单纯优化 CTQ 集的分类性能，算法得到很高的交叉验证精度，但是相对本文算法，其预测精度相对交叉验证精度出现了更明显的下降。这说明 Memetic 算法在识别 CTQ 时出现了更严重的过拟合问题，使得算法选择了过多无关、冗余质量特性，对算法预测精度造成影响，最终影响 CTQ 识别效果。IG 算法在 CTQ 识别时，单独评估每个质量特性影响产品质量的程度，没有考虑到质量特性间的冗余性，不能有效过滤冗余质量特性，所以识别的质量特性数也较多。总体来看，本文算法有更好的 CTQ 识别结果。综上，本文算法进行 CTQ 识别时能够快速有效收敛，并能有效过滤无关、冗余质量特性，识别复杂产品 CTQ。

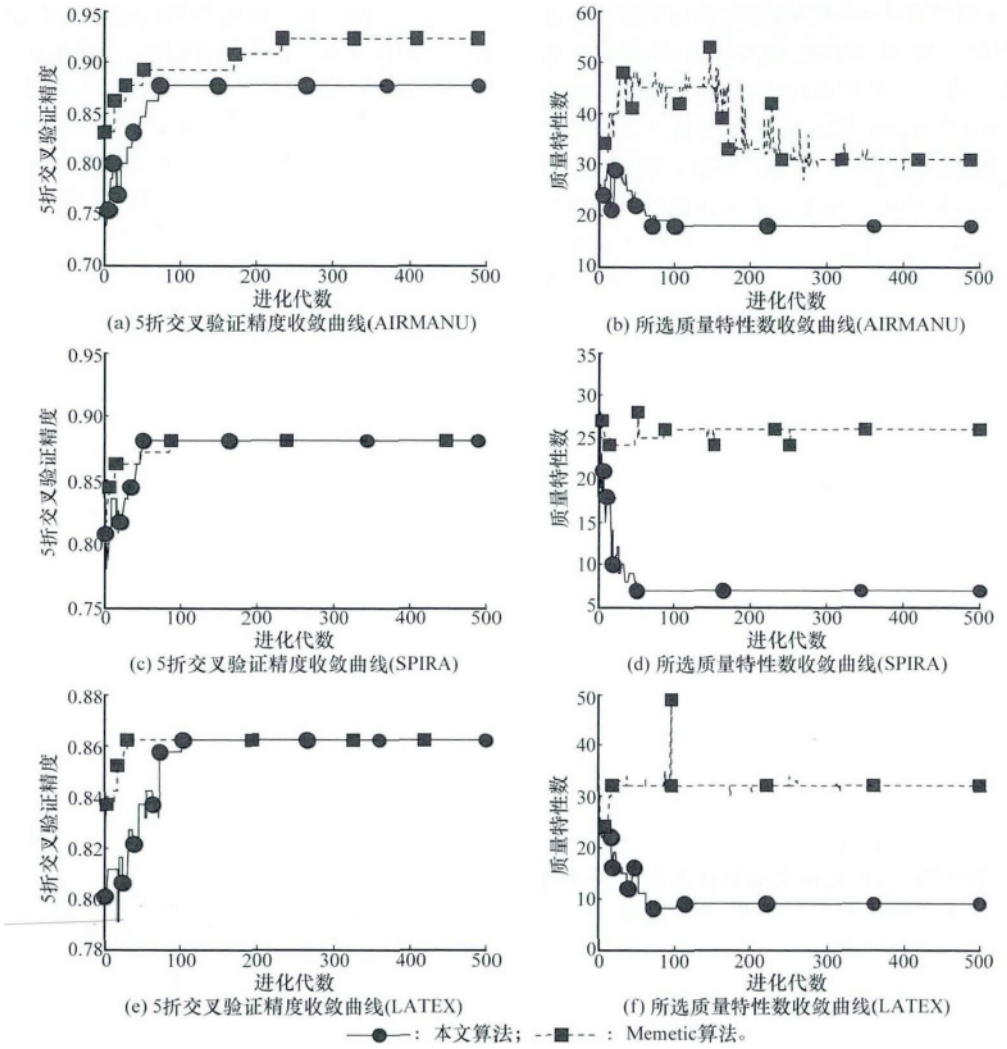

图 2 收敛性能图

表 3 各算法 CTQ 识别结果

数据集	本文算法		IG 算法		Memetic 算法	
	预测精度/%	质量特性数	预测精度/%	质量特性数	预测精度/%	质量特性数
AIRMANU	86.36	18	77.27	40	77.27	31
SPIRA	86.11	7	77.78	20	75.00	26
LATEX	84.85	9	77.27	27	84.85	32
平均	85.77	11.33	77.44	29.00	79.04	29.67

4 结 论

复杂产品包含大量质量特性，为了有效识别影响产品质量的 CTQ，本文建立基于 GSA 算法特征选择算法，并用于 CTQ 识别。首先，所提算法将 GA 算法与 SA 算法融合，兼有两者的优点，有不错全局搜索与局部搜索能力。其次，为了处理对训练集的过拟合问题，提高算法过滤无关、冗余质量特性的能力，本文提出综合的适应度函数应用于所提算法。所提综合适应度函数，能够使算法在优化过程中同时最大化 CTQ 集分类性能和最小化 CTQ 集所选质量特性数。本文通过算例验证了算法的有效性，结果表明所提算法能够快速有效收敛，同时算法在识别更少的 CTQ 的同时得到更高预测精度。说明算法能够有效过滤无关、冗余质量特性，并有效识别 CTQ。如何将所提算法扩展到不平衡数据的 CTQ 识别，是今后需要做的工作。

参考文献：

[1] Li B H. Key technologies in informatization of complex product: complex product integrated manufacturing system[J]. *Manufacture Information Engineering of China*, 2006(14): 19-23. (李伯虎. 复杂产品制造信息化的重要技术——复杂产品集成制造系统[J]. 中国制造业信息化, 2006(14): 19-23.)

[2] Thornton A C. A mathematical framework for the key characteristic process[J]. *Research in Engineering Design*, 1999, 11(3): 145-157.

[3] Jia G Z, Bai M. An approach for manufacturing strategy development based on fuzzy-QFD[J]. *Computers & Industrial Engineering*, 2011, 60(3): 445-454.

[4] Lee D J, Thornton A C. The identification and use of key characteristics in the product development process[C]// *Proc. of the ASME Design Engineering Technical Conferences and Computers in Engineering Conference*, 1996: 211-217.

[5] Yan W, He Z, Tian W M, et al. Research on complex products critical-to-quality characteristics identification method based on IG[J]. *Industrial Engineering and Management*, 2012, 17(1): 55-60.(**闫伟，何桢，田文萌，等. 基于** IG **的复杂产品关键质量特性识别方法的研究**[J]. **工业工程与管理**，2012，17(1)：55-60.)

[6] Hua J, Tembe W D, Dougherty E R. Performance of feature-selection methods in the classification of high-dimension data[J]. *Pattern Recognition*, 2009, 42(3): 409-424.

[7] Yao X, Wang X D, Zhang Y X, et al. Ensemble feature selection algorithm based on Markov blanket and mutual information[J]. *Systems Engineering and Electronics*, 2012, 34(5): 1046-1050.(**姚旭，王晓丹，张玉玺，等. 基于** Markov blanket **和互信息的集成特征选择算法**[J]. **系统工程与电子技术**，2012，34(5)：1046-1050.)

[8] 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.(**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]. **系统工程理论实践**，2014，34(5)：1230-1236.)

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

[10] Pudil P, Novovičová J, kittler J. Floating search methods in feature selection[J]. *Pattern Recognition Letters*, 1994, 15(11): 1119-1125.

[11] Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection[J]. *Journal of Chemometrics*, 1992, 6(5): 267-281.

[12] Vasighi M, Zahraei A, Bagheri S, et al. Diagnosis of coronary heart disease based on 1H NMR spectra of human blood plasma using genetic algorithm based feature selection[J]. *Journal of Chemometrics*, 2013, 27(10): 318-322.

[13] Min S H, Lee J, Han I. Hybrid genetic algorithms and support vector machines for bankruptcy prediction[J]. *Expert Systems with Applications*, 2006, 31(3): 652-660.

[14] 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.

[15] Pacheco J, Casado S, Núñez L, et al. Analysis of new variable selection methods for discriminant analysis[J]. *Computational Statistics & Data Analysis*, 2006, 51(3): 1463-1478.

[16] Li J H, Yu F, Fan F J. Ship block assembly selquence optimization based on genetic simulated annealing algorithm[J]. *Computer Integrated Manufacturing Systems*, 2013, 19(1): 39-45.(**李敬花，余峰，樊付见. 基于遗传模拟退火融合算法的船舶分段装配序列优化**[J]. **计算机集成制造系统**，2013，19(1)：39-45.)

[17] He X L, Bi Y M. Modeling and optimization of formation air-to-ground attack fire distribution based on simulated annealing genetic algorithm[J]. *Systems Engineering and Electronics*, 2014, 36(5): 900-904.(**贺小亮，毕义明. 基于模拟退火遗传算法的编队对地攻击火力分配建模与优化**[J]. **系统工程与电子技术**，2014，36(5)：900-904.)

[18] Oh I S, Lee J S, Moon B R. Hybrid genetic algorithms for feature selection[J]. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2004, 26(11): 1424-1437.

[19] 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.

[20] Bermejo P, Gámez J A, Puerta J M. Speeding up incremental wrapper feature subset selection with Naive Bayes classifier[J]. *Knowledge-Based Systems*, 2014, 55: 140-147.

[21] 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.

作者简介：

李岸达(1989-)，**男，博士研究生，主要研究方向为质量工程、智能算法。**
E-mail：lianda1989@gmail.com

何　桢(1967-)，**男，教授，博士，主要研究方向为质量管理、质量工程。**
E-mail：zhhe@tju.edu.cn

何曙光(1979-)，**男，教授，博士，主要研究方向为质量工程、信息系统。**
E-mail：shuguanghe@tju.edu.cn