文章编号:1007-5429(2014)03-0053-07

基于 Filter 与 Wrapper 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:为了解决复杂产品关键质量特性(CTQs)识别问题,提出基于过滤(Filter)算法与包裹(Wrapper)算法的改进混合特征选择算法。首先应用 Filter 算法对复杂产品质量特性进行排序,接着应用 Wrapper 算法识别关键质量特性。提出一种新的方法 FNO,确定 Wrapper 阶段所选关键质量特性数。应用过滤算法 ReliefF 构建混合算法 ReliefF-W。算例分析表明,ReliefF-W 能够有效进行 CTQ 识别。与一种传统混合算法相比,ReliefF-W 能够在保证学习算法有较好预测精度的同时,识别出更少关键质量特性。

关键词:关键质量特性;复杂产品;特征选择;过滤算法;包裹算法;ReliefF
中图分类号:F406.3 **文献标识码**:A

Identification of Critical-to-quality Characteristics for Complex Products Based on Bybrid Algorithm of Filter and Wrapper

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), a hybrid algorithm of Filter and Wrapper is proposed. Quality characteristics are ranked by Filter algorithm firstly, and then Wrapper algorithm is applied to select the CTQs. A new method named FNO is proposed to calculate the best number of CTQs during the Wrapper phase. A Filter algorithm, ReliefF, is introduced to structure the hybrid algorithm named ReliefF-W. Experimental result illustrates that ReliefF-W is efficient in CTQ identification. Compared with a traditional hybrid algorithm, ReliefF-W can select fewer CTQs while ensuring high prediction accuracy.

Key words: critical-to-quality characteristics(CTQs); complex products; feature selection; filter algorithm; wrapper algorithm; ReliefF

1 引言

复杂产品是指具有“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[1]。从质量角度来看,产品结构越复杂,构成零部件数量越多,对其进行质量控制的难度越大,出现质量缺陷的概率也就越大。由于影响复杂产品的零件数量多、工序多,每个零件和每道工序都有多个质量特性,这些质量特性对最终产品的影响关系也非常复

收稿日期:2013-06-20; 修回日期:2013-10-20
基金项目:国家杰出青年科学基金资助项目(71225006);国家自然科学基金资助项目(70931004)
作者简介:李岸达(1989-),甘肃渭源人,博士研究生,主要研究方向为质量工程。

杂，难以用现有的工程知识解释或揭示这种影响关系。如在某复杂机电产品的装配过程中，尽管所有的零件均符合公差要求，但是有的产品装配后表现为合格，有的则不合格，很难解释哪个零件或哪道工序是关键影响因素，只能将所有零件的公差范围收缩，导致加工成本增加。因此，从众多的零件和工序质量特性中识别出关键质量特性（Critical-to-quality Characteristics，CTQs）对复杂产品的质量控制具有非常重要的意义。从根本上来说 CTQ 识别是降维的过程。数据挖掘领域、机器学习领域的特征选择（Feature Selection）或变量选择（Variable Selection）[2] 可以用于复杂产品 CTQ 识别。

2　关键质量特性识别研究综述

通常确定产品质量特性时，首先要对产品特性逐层分解，由上到下分解为产品特性、部件特性、零件特性、工艺特性等，这一过程称为关键特性展开（Key Characteristics Flowdown）[3]，在层级展开之后结合一些定性、定量的方法进一步识别关键特性，此类研究成果见文献[4-6]。另一类 CTQ 识别方法是质量功能展开（Quality Function Deployment，QFD），该方法体现了市场导向，通过对顾客需求的逐层展开确定产品的关键质量特性（各级零部件特性和过程特性）[7]。但是，复杂产品结构复杂，进行关键特性展开后会得到高维度的质量特性，各质量特性间的关系错综复杂，对每一个质量特性进行定性、定量分析将产生大量人力、时间成本。质量特性的急剧增加使得传统方法失去效果[8]。在高纬度带来的负面影响下，QFD 的质量矩阵会变得难以确定，导致 QFD 不能有效进行 CTQ 识别[9]。

数据挖掘领域、机器学习领域的特征选择（Feature Selection）方法，可以有效用于处理高维数据[10]。CTQ 识别问题可以抽象成一个特征选择问题。其中，产品“质量特性”对应于特征选择算法中的“特征”，“产品分级”对应于“类标签”。特征选择算法通过分析类标签与各个特征之间的关系，对特征进行筛选，将原来的大的特征集合降维成一个小的特征集合。将特征选择算法用于 CTQ 识别时，通过不同特征选择算法分析“产品分级”与各个“质量特性”的关系，筛选出最有效的质量特性子集。

特征选择可以分为过滤（Filter）算法、包裹（Wrapper）算法、和嵌入（Embedded）算法。Filter 算法是学习算法之前的一个预处理步骤，它在学习之前过滤掉不相关的特征[11]，这类算法的优点是计算速度快，输出结果可以用于不同的学习算法。信息增益（IG）算法、Relief 以及其改进 ReliefF 都属于过滤算法[12,13]。Wrapper[14] 算法在选择特征子集时将学习算法本身作为评估函数的一部分，学习算法分类性能是评估特征重要程度的评价标准。这类算法的优点是对于特定学习算法，其分类准确率较高。缺点是时间复杂度较高，随着维度的增高其计算时间急剧增加，这种现象也称为维度灾难（Curse of Dimensionality）。另一类是 Embedded 算法，最优特征子集的选择被集成在分类器构建的过程中，可以认为学习过程和特征选择过程是同时进行的，最常见的 Embedded 算法是决策树，例如 C4.5[15]。

文献[9]将 IG、ReliefF 用于复杂产品 CTQ 识别中，并通过朴素贝叶斯分类器验证了方法的有效性。文献[9]使用了设定阈值 δ 的方式筛选特征（权重小于 δ 的特征被过滤掉），但是 δ 的选取是一个难点，不同的数据集选取合理的 δ 会有很大不同，另一方面该方法在选取特征子集时没有考虑学习算法的分类精度。文献[16]在将特征选择算法应用于高维基因组微阵列数据（Genomic Microarray Data）时，提出一种混合特征选择算法。首先应用 Filter 算法筛选特征，将特征数由原来的几千个筛选到几百个，接着应用 Wrapper 算法选择特征子集。混合算法兼有 Filter 算法与 Wrapper 算法两者的优点：首先，相对于传统 Wrapper 算法，它有更小的计算复杂度。其次，相对于 Filter 算法，在筛选特征子集的时候考虑了学习算法的分类性能。但是文献[16]的算法应用于复杂产品 CTQ 识别会存在以下问题。第一，该算法最终选取的特征（质量特性）过多，不能保证有效识别关键质量特性。第二，文献[16]没有说明 Wrapper 算法阶段特征的选择顺序，不同的选择顺序得到不同特征（质量特性）子集，不能得到唯一结果，不能将其直接用于 CTQ 识别。为了解决文献[9]及文献[16]存在的问题，本文提出改进的混合特征选择算法，并对该混合算法进行性能分析，验证其有效性。

3　特征过滤算法 ReliefF

ReliefF 算法是一个经典的 Filter 算法，它通过一定的策略分析各特征与类标签的关系来判断各特征的重要性，最终该算法得到各特征的权重，权重值越大说明特征越重要。该算法是特征过滤算法 Relief 算法的改进。Relief 算法最早由 Kira 与 Rendell[17,18] 于 1992 年提出，该算法只能解决类标

签是两类的分类问题，不能处理缺失值，并且对噪声较敏感。Kononenko 随后提出了 Relief 算法的改进 ReliefF，该算法解决了 Relief 存在的以上问题[13]。

令训练集为 Ω，其特征集合为$\{F_1,F_2,\cdots,F_L\}$，类标签集合为$\{C_1,C_2,\cdots,C_n\}$，则每个样本实例 X_i 是一个 L 维的向量$(X_{i,1},X_{i,2},\cdots,X_{i,L})$则 ReliefF 的基本步骤可以描述为：首先，从 Ω 中选取一个样本实例 X_i（对应的类标签为 $class(X_i)$），然后从 Ω 中找到离 X_i 最近的 K 个同类样本实例 $H_{ij}(j\in(1,\cdots,K))$组成 X_i 的同类最近集 NH。其次，从每一类 $C_p(C_p\neq class(X_i))$中选取 K 个与 X_i 不同类的最近样本实例$M_{ij}(C_p)(j\in(1,\cdots,K))$组成 X_i 的异类最近集 $NM_p,(p\in(1,\cdots,n))$。最后，计算各特征 $F_l(l\in(1,\cdots,L))$的权重 $W(F_l)$，其取值依赖 F_l 区分 NH 与 NM_p 中实例的能力，区分能力越强 $W(F_l)$取值越高。最终的属性权重 $W(F_l)$是挑选 M 个 $X_i(i\in(1,\cdots,M))$之后平均得到的结果。$W(F_l)$的计算公式如式(1)所示。

$$W(F_l)=\sum_{i=1}^{M}(-\sum_{j=1}^{K}diff(F_l,X_i,H_{ij})/(M\cdot K)+\sum_{c\neq class(X_i)}(\frac{p(c)}{1-class(X_i)}\sum_{j=1}^{K}diff(F_l,X_i,M_{ij}(c)))/(M\cdot K)) \quad (1)$$

式(1)中，$p(c)$代表类 c 的概率，可以通过样本实例进行估计。$diff(F_l,X_i,X_j)$是区分度函数，该函数计算特征 F_l 区分样本实例 X_i 和 X_j 的能力。当 F_l 是属性(nominal)变量时，$diff(F_l,X_i,X_j)$定义如下：

$$diff(F_l,X_i,X_j)=\begin{cases}0;X_{i,l}=X_{j,l}\\1;X_{i,l}\neq X_{j,l}\end{cases} \quad (2)$$

当 F_l 是数值(numerical)变量时，$diff(F_l,X_i,X_j)$定义如下：

$$diff(F_l,X_i,X_j)=\frac{|X_{i,l}-X_{j,l}|}{\max(F_l)-\min(F_l)} \quad (3)$$

其中 $\max(F_l)$和 $\min(F_l)$分别表示特征 F_l 的最大取值和最小取值。

ReliefF 最终得到各特征 F_l 的权重 $W(F_l)$，权重越高该特征越重要，根据权重的大小可以对各特征进行排序。将 ReliefF 算法用于 CTQ 识别时，各特征 F_l 对应于质量特性，类标签对应于产品的质量水平，通过该算法能够得到各质量特性的权重，根据权重大小可以判断质量特性的重要程度。

4 构建混合特征选择算法

4.1 传统特征选择算法的不足

使用 ReliefF 算法能够根据权重大小对特征(质量特性)进行排序，越靠前的特征越重要。但是选择多少个特征(质量特性)需要人为决定。通常有两种思路，一种是选择前 n 个特征(质量特性)，参数 n 人为决定[11]；另一个思路是确定一个阈值 δ，保留权重大于 δ 特征(质量特性)[17]。但是这两种思路的缺点也是明显的，参数 n 以及阈值 δ 的取值不易确定。文献[16]在进行高维基因组微阵列数据(Genomic Microarray Data)中的特征选取时应用了一种新的思路确定选择的特征数，其核心思想如下：假定数据集有 L 个特征。每次选择 m(从 1 到 L)个特征，并应用学习算法得到对应的交叉验证精度 $Acc(m)$。最后选择最高交叉验证精度 $Acc(M)$对应的特征数 M。将文献[16]确定特征数的方法应用于确定 IG 与 ReliefF 需要保留的特征数，能够解决以上两种思路不易确定参数的问题，因为这种方法应用了分类器的分类性能作为保留特征数的依据。由于使用了分类器性能作为判断依据，所以文献[16]算法是一种 Wrapper 算法。

但是直接应用文献[16]算法确定 ReliefF 需要保留的特征(质量特性)数，也存在问题，这种方法会保留过多的特征(质量特性)。两组数据的“特征数——交叉验证精度图”如图 1 所示。横轴代表特征数，竖轴代表交叉验证精度。横坐标是离散的，即特征数只能取整数。采用文献[16]的算法所得结果在图中用圆圈表示。图 1-a 中，原始特征有 120 个，使用文献[16]算法最终选择特征数为 35(原始特征数的 29%)，特征选择有不错效果。图 1-b 中，原始特征数为 35，使用文献[16]算法最终选择的特征数为 32(原始特征数的 91%)，特征选择的效果并不明显。从这两个实际数据的例子可以看到，文献[16]算法在有些情况能够比较有效选取合理的特征子集，但在另一些情况下这种算法效果不够理想。

图 1-b 中，若选取特征数为 $SNum$(圆点处)，特征数明显减少，对应的交叉验证精度略小于最高交叉验证精度，这说明选取特征数为 $SNum$ 能够得到理想结果。从特征数 $SNum$ 之后每增加一个特征其交叉验证精度的增加已经不够明显，之后的特征对分类的贡献已经不大。对于 CTQ 识别，其目的是识别最重要的质量特性，能够有效从众多质量特性中识别最重要的少数质量特性是首要目标。若

CTQ 识别时出现图 1-b 中的情况，选取 $SNum$ 个特征（质量特性）其预测精度已经足够高，选取 $SNum$ 个特征（质量特性）是合理的。另一方面，选取更多的特征（质量特性）并不能保证使用测试集得到的预测精度也足够高，因为存在过拟合的可能性。

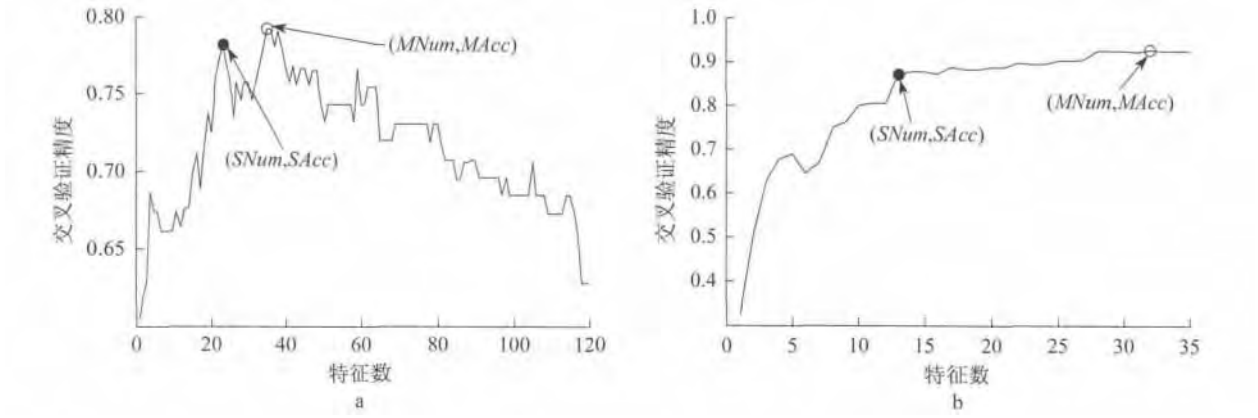

图 1　特征数——交叉验证精度图

4.2　选择特征数的确定

由 4.1 节的分析可知，对于 CTQ 识别，最优的质量特性数并不一定是最高交叉验证精度对应的质量特性数。理想的结果是选择一个合理的质量特性子集，其对应的交叉验证分类精度相对最高值没有显著下降，而质量特性数有比较明显的降低。可以将其描述为两个目标：第一，比较高的交叉验证精度；第二，比较少的质量特性（特征）数。

图 1-b 中，最大特征数用 N 表示，$Acc(k)$ 表示特征数 $k(k\in\{1,2,\cdots,N\})$ 对应的交叉验证精度，则图中的点可以用 $(k,Acc(k))$ 表示。最高交叉验证精度用 $MAcc$ 表示，对应的特征数为 $MNum$，则在图中的坐标为 $(MNum,MAcc)$，即使用文献[16]算法得到的结果。由以上分析，选取圆点 $(SNum,SAcc)$ 对应的特征数比较理想。从点 $(1,Acc(1))$ 到点 $(MNum\text{-}1,Acc(MNum\text{-}1))$，每一个点都能与点 $(MNum,MAcc)$ 连成一条直线，最终可以连成 $MNum\text{-}1$ 条这样的直线，每一条直线的斜率表示为 $slope(k)$。可以看到，若 $k<SNum$，则斜率 $slope(k)$ 较大，$k>SNum$ 则斜率 $slope(k)$ 明显减少。这些直线斜率的大小可以作为寻找最优特征数的依据。若设定一个参数 λ，从第一个点开始计算，计算各点的 $slope(k)$，找到第一个使得 $slope(K)\leqslant\lambda$ 的值 K，则可得到 $SNum=K$，这样就能找到一个合理的特征数。

基于以上分析，本文提出一种新的算法选择最优特征数，用函数的形式表示，命名该函数为 FNO，定义如表 1。函数 FNO 的输入是交叉验证精度的集合 $ASet=\{Acc(k)\mid k\in\{1,2,\cdots,N\}\}$，以及参数 τ；输出是该算法得到的最优特征数 $FNum$。表 1 中 $\lambda=\tau/(100\cdot MNum)$，这样处理可以用一个更为直观的参数 τ 作为输入参数，便于进行参数设置与解释。设置参数 τ，就是允许交叉验证精度相对最高值下降量小于 $\tau\%$。若最终选取的特征数为 $l(0<l<MNum)$，则有(4)式推导。

$$\begin{aligned}&slope(l)\leqslant\lambda\\&\Rightarrow(MAcc-Acc(l))/(MNum-l)\leqslant\tau/(100\cdot MNum)\\&\Rightarrow(MAcc-Acc(l))\leqslant(\tau/(100\cdot MNum))\cdot(MNum-l)\\&\Rightarrow(MAcc-Acc(l))<\tau/100\end{aligned}\tag{4}$$

从以上推导中可以看到最终选取的特征数 l 对应的交叉验证精度低于最高值的量小于 $\tau\%$。τ 值选取越大，可以选到的特征数就越少。理论上 $\tau\in[0,100]$，τ 取 0 时所得结果就是文献[16]结果，τ 取 100 时选取特征数 $l=0$，实际上 τ 的选取可以远远小于 100（见算例分析），通过调节 τ 的大小，可以控制所选特征子集的大小。

对图 1-a 中的数据集使用 FNO 函数选择最优特征数，令 τ 为 5，得到 $FNum=SNum$，最终选择的点为 $(SNum,SAcc)$，相对于文献[16]算法所选点 $(MNum,MAcc)$，该点特征数有较明显降低，交叉验证精度只有小幅下降。对于图 1-b 中的数据集，令 τ 为 10，得到 $FNum=SNum$，所选点为 $(SNum,SAcc)$，若令 τ 为 15，最终所选点仍然为 $(SNum,SAcc)$，可以看到 FNO 函数是比较稳定的，只要将 τ 选取在一个合理的范围，就能达到比较理想的效果。若 τ 取 0，则得到的 $FNum=MNum$，即 $\tau=0$ 时，

FNO函数找到得到的结果与文献[16]算法相同，本质上FNO函数是在文献[16]算法基础上的扩展。

表1　FNO函数伪代码

```
Input：集合 ASet={Acc(k)|k∈{1,2,…,N}}，参数 τ
Output：特征数 FNum
function　FNum=FNO(ASet,τ)
    从集合 ASet 中找到 MNum 和 MAcc；
    处理参数 τ，令 λ=τ/(100·MNum)；
    for i=1 to MNum-1
      slope(i)=(MAcc-Acc(i))/(MNum-i)
      if slope(i)<=λ
        FNum=i；
        return；
      end if
    end for
end function
```

4.3　构建混合特征选择算法

基于4.1节与4.2节的分析，本节提出一种改进的混合算法解决CTQ识别问题。算法思路如下：首先，应用Filter算法（如ReliefF）对数据集进行特征（质量特性）排序，得到经过排序的特征（质量特性）集 FR。其次，应用Wrapper算法选择最优特征（质量特性）子集：每次从 FR 中选择前 i 个特征（质量特性），使用这 i 个特征（质量特性）应用学习算法进行分类，得到对应的交叉验证精度，直到 FR 中最后一个特征（质量特性）被选中，从而得到该数据集的“特征数——交叉验证精度图”。使用本文提出的 FNO 函数得到最优特征（质量特性）数 $FNum$。特征集合 FR 中前 $FNum$ 个特征（质量特性）组成最终所选特征（质量特性）子集。为了验证算法的有效性，需要将数据集分为训练集和测试集，使用训练集得到特征（质量特性）子集，使用测试集对结果进行验证，算法处理框架如图2所示。算法步骤如下：

Input　数据集Data（其特征集为 FS，包含特征数为 N），调节参数 τ；

Output　最终特征子集 FR-$FNum$，预测精度 $TestAcc$；

Step1　将数据集Data，分为训练集Trainset和测试集Testset；

Step2　使用能够特征排序的Filter算法处理Trainset，得到每个特征 F_i 权重 $W(i)$，根据 $W(i)$，按重要程度将 F_i 排序，得到排序后特征集合 FR；

Step3　令 $i=1$；

Step4　选取 FR 中前 i 个特征，组成特征集 FR-i；

Step5　使用集合 FR-i 中的特征，应用训练集Trainset进行5折交叉验证，得到交叉验证精度 $Acc(i)$；

Step6　$i=i+1$，如果 $i<N+1$，返回Step4，否则，下一步；

Step7　令 $ASet=\{Acc(k)|\ k\in\{1,2,\cdots,N\}\}$；

Step8　$FNum=\mathrm{FNO}(ASet,\tau)$　（表1所示）；

Step9　FR-$FNum$ 为最终选取的特征子集。使用Trainset训练分类器，应用Testset进行预测，得到 FR-$FNum$ 对应的预测精度 $TestAcc$。

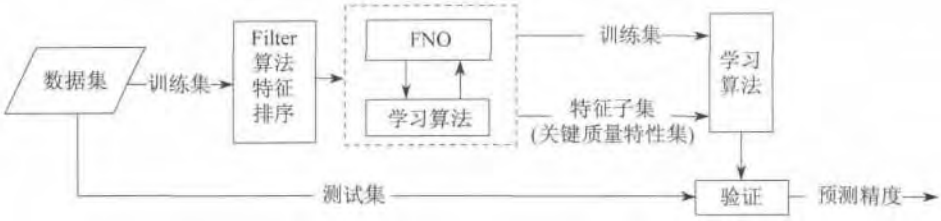

图2　Filter与Wrapper混合算法特征选择框架

5　算例分析

5.1　实验条件

算例集选自UCI数据库SECOM数据集[19]，闫伟[9,20]等人应用该数据集进行CTQ识别分析。数据集共有样本1564个，每个样本包含590个质量特性，令质量特性标号为“No.1”到“No.590”；样本分为合格产品和不合格产品2类，合格产品样本数量为1463个，不合格产品样本数量为104。该数据集是一种严重不平衡数据集，不平衡性严重影响特征选择算法的CTQ识别效果[20]，所以在进行CTQ识别之前需要先将数据作平衡化处理。本文使用一种经典的处理策略——欠抽样（Under-sampling），随机从合格产品中挑选104个样本，与原不合格产品样本组成新的数据集，命名该数据集为US-SECOM，其信息如表2所示。

分别应用文献[16]算法以及本文提出的混合特征选择算法对数据集进行CTQ识别，两种算法中

都要用到 Filter 算法进行特征排序，本实验使用 ReliefF 算法构建混合算法（注：文献[16]算法原本没有排序这一阶段，为了便于与本文算法比较，对文献[16]算法增加排序步骤）。用 ReliefF-T 表示文献[16]算法，用 ReliefF-W 表示本文混合算法。在使用两种算法进行 CTQ 识别时，特征选择算法中的特征对应于质量特性，类标签对应于产品质量水平（合格/不合格）。特征选择算法中要用到学习算法（分类器），这里选用一种经典的学习算法灵活朴素贝叶斯分类器（Flexible Naive Bayes）[21]。进行算法验证时同样需要学习算法，同样选用灵活朴素贝叶斯分类器。为了验证算法的有效性，将数据集 US-SECOM 分为两部分，分别是训练集以及测试集，训练集包含样本 166 个，测试集包含样本 44 个，使用训练集得到关键质量特性集合，再应用训练集验证结论的有效性，实验流程如图 2 所示。实验结果将从预测精度以及降维水平两个方面进行评价[22]。

表 2　US-SECOM 数据集信息

数据集	质量特性数	样本	合格/不合格	分类
US-SECOM	590	208	104/104	2

5.2　实验结果

表 3 所示为采用 ReliefF-T 与 ReliefF-W 对数据集 US-SECOM 的 CTQ 识别结果。表中 org 行表示未经 CTQ 识别的结果。在应用 ReliefF-W 算法时选取参数 $\tau=1$ 到 $\tau=7$ 得到识别结果。从实验结果中可以看到，使用 ReliefF-T 得到的 CTQ 集合包含 77 个质量特性，占总质量特性数的 13.05%，相对于原始数据集（Org 行），预测精度从 61.90% 提高到 64.29%。ReliefF-W 得到的结果在 $\tau=1$ 时结果与 ReliefF-T 得到的结果相同。在 $\tau=2,3,4$ 时 ReliefF-W 得到的结果相同，预测精度为 71.43%，相对 ReliefF-T 与原始预测精度有比较明显提高，识别的 CTQ 数量为 41 个，占总数的 6.95%。当 $\tau=5,6,7$ 时，ReliefF-W 识别的 CTQ 数分别是 7、5、5，但是预测精度相对原始预测精度有所下降。从预测精度来看，当 $\tau=2,3,4$ 时 ReliefF-W 的预测精度达到最高，相对 ReliefF-T 提高了 7.14%。从降维水平来看，当 $\tau=2,3,4$ 时，ReliefF-W 识别的 CTQ 数量占 ReliefF-T 识别数量的 53.25%。从降维水平和预测精度综合考虑，$\tau=2,3,4$ 时，ReliefF-W 得到的识别结果明显好于 ReliefF-T。当 τ 取较大值 5、6、7 时，ReliefF-W 能够识别更少 CTQ，但是预测精度也出现下降，结果没有 $\tau=2,3,4$ 时好。随着 τ 的增大，识别的 CTQ 数是下降的，这与理论推导是一致的。

表 3　采用各算法得到的 CTQ 识别结果

算法名称		所得 CTQ 集	CTQ 数（占总数百分比）	预测精度
Org		No. 1；No. 2；…；No. 590	590(100%)	61.90%
ReliefF-T		No. 77；No. 81；No. 60；No. 29；No. 512；No. 1；No. 356；No. 218；No. 320；No. 324；No. 38；No. 13；No. 41；No. 125；No. 19；No. 478；No. 461；No. 456；No. 583；No. 460；No. 206；No. 188；No. 317；No. 184；No. 407；No. 32；No. 122；No. 203；No. 269；No. 541；No. 113；No. 201；No. 130；No. 342；No. 406；No. 490；No. 318；No. 124；No. 153；No. 61；No. 288；No. 424；No. 488；No. 434；No. 268；No. 426；No. 572；No. 354；No. 216；No. 166；No. 540；No. 339；No. 66；No. 44；No. 492；No. 220；No. 22；No. 65；No. 133；No. 286；No. 144；No. 417；No. 108；No. 151；No. 162；No. 165；No. 3；No. 28；No. 297；No. 18；No. 202；No. 570；No. 301；No. 80；No. 358；No. 475；No. 280	77(13.05%)	64.29%
ReliefF-W	$\tau=1$	同上	77(13.05%)	64.29%
	$\tau=2$	No. 77；No. 81；No. 60；No. 29；No. 512；No. 1；No. 356；No. 218；No. 320；No. 324；No. 38；No. 13；No. 41；No. 125；No. 19；No. 478；No. 461；No. 456；No. 583；No. 460；No. 206；No. 188；No. 317；No. 184；No. 407；No. 32；No. 122；No. 203；No. 269；No. 541；No. 113；No. 201；No. 130；No. 342；No. 406；No. 490；No. 318；No. 124；No. 153；No. 61；No. 288	41(6.95%)	71.43%
	$\tau=3$	同上	41(6.95%)	71.43%
	$\tau=4$	同上	41(6.95%)	71.43%
	$\tau=5$	No. 77；No. 81；No. 60；No. 29；No. 512；No. 1；No. 356	7(1.19%)	54.76%
	$\tau=6$	No. 77；No. 81；No. 60；No. 29；No. 512	5(0.85%)	59.52%
	$\tau=7$	No. 77；No. 81；No. 60；No. 29；No. 512	5(0.85)	59.52%

从以上的结果可以看到，当 $\tau=2,3,4$ 时，ReliefF-W 的识别结果要明显好于 ReliefF-T 的识别结果，ReliefF-W 能够在不降低预测精度的同时识别更少 CTQ。对 ReliefF-W，$\tau=2,3,4$ 是合理的 τ 的取值，该取值并不是一个确定的值，它是在一个区间内的，这一特性便于选择合理的 τ 值。综上所述，本文提出的混合特征选择算法 ReliefF-W 要优于文献[16]算法 ReliefF-T。

6 结论

本文在一种传统的 Filter 与 Wrapper 混合算法的基础上，提出改进的 Filter 与 Wrapper 混合算法，并选取一种经典的 Filter 算法 ReliefF 构建混合算法 ReliefF-W，并将其用于复杂产品 CTQ 识别。算例分析表明 ReliefF-W 能够有效进行 CTQ 识别。与文献[16]算法相比，ReliefF-W 能够在不降低预测精度的前提下识别更少的 CTQ。总体来说，本文混合算法 ReliefF-W 有更高 CTQ 识别效率，本文算法要优于文献[16]算法。本文选用了 ReliefF 算法构建混合算法，选取更多的 Filter 算法构建混合算法，并从中寻找一种或几种有效处理 CTQ 识别的算法是今后需要做的工作，此外对比不同学习算法的 CTQ 识别结果也是很有必要的。

参考文献：

[1] 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] Guyon I，Elisseeff A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research，2003，3：1157-1182.

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

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

[5] 刘志存，邹冀华，范玉青. 飞机制造中关键特性的定义与管理[J]. 计算机集成制造系统，2007，13(10)：2013-2018.

[6] 魏丽，郑联语. 概要工艺规划中关键特性的识别过程及方法[J]. 计算机集成制造系统，2007，13(1)：147-152.

[7] 马林，何桢. 六西格玛管理[M]. 北京：中国人民大学出版社，2007.

[8] Pierre E S，Tuv E. Robust，non-redundant feature selection for yield analysis in semiconductor manufacturing[J]. Advances in Data Mining. Applications and Theoretical Aspects，2011，6870：204-217.

[9] 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.

[10] 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.

[11] Blum A L，Langley P. Selection of relevant features and examples in machine learning[J]. Artificial Intelligence，1997，97(1)：245-271.

[12] Oaksford M，Chater N. Information gain explains relevance which explains the selection task[J]. Cognition，1995，57：97-108.

[13] Kononenko I. Estimating attributes：analysis and extensions of RELIEF[J]. Machine Learning：ECML-94，1994，784：171-182.

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

[15] Quinlan J R. C4. 5：Programs for Machine Learning[M]. San Mateo，CA：Morgan Kaufmann，1993.

[16] Xing E P，Jordan M I，Karp R M. Feature selection for high-dimensional genomic microarray data[C]. International Conference on Machine Learning. San Francisco，CA：Morgan Kaufmann，2001：601-608.

[17] Kira K，Rendell L A. The feature selection problem：Traditional methods and a new algorithm[C]. Proceedings of the Tenth National Conference on Artificial Intelligence. Cambridge，MA：MIT Press，1992：129-134.

[18] Kira K，Rednell L A. A practical approach to feature selection[C]. Proceedings of the ninth international workshop on Machine learning. Los Altos，CA：Morgan Kaufmann，1992：249-256.

[19] Frank A，Asuncion A.（2010）. UCI Machine Learning Repository[http://archive. ics. uci. edu/ml]. Irvine，CA：University of California，School of Information and Computer Science.

[20] 闫伟，何桢，田文萌，等. 基于 EM 的不平衡数据关键质量特性识别[J]. 工业工程与管理，2012，17(4)：38-42.

[21] John G H，Langley P. Estimating Continuous Distributions in Bayesian Classifiers[C]. Eleventh Conference on Uncertainty in Artificial Intelligence. San Francisco，CA：Morgan Kaufmann 1995：338-345.

[22] Witten I H，Frank E，Hall M A. Data Mining：Practical Machine Learning Tools and Techniques. 3rd Ed[M]. Burlington，MA：Morgan Kaufmann，2011.