基于遗传算法与最大最小原理的故障模式特征选择*

谢涛 张育林

(国防科技大学航天技术系 长沙 410073)

摘 要 在诸如液体火箭发动机等复杂动力学系统的故障诊断中, 监控参数组的优选问题一直受到工程技术人员的高度重视. 本文提出了综合样本矢量方向离散度概念, 以此作为故障特征参数的优选准则; 然后利用经过改进的遗传算法, 对某液体火箭发动机常见故障的诊断进行了特征参数组的优选. 在改进的遗传算法中, 采用了非常简洁而高效的染色体编码, 针对特征优选的组合优化类问题专门设计了一种特殊的基因迁移算子, 并引进了父本个体适应值的动态调整技术与共享函数. 数值实验结果表明, 该算法具有理想的效果.

关键词 统计聚类、样本矢量方向离散度、故障特征参数选择、故障仿真、优化、遗传算法
分类号 V 433

Max-Min Principle Based-Selection for the Optimal Feature Parameters in Fault Diagnosis Using Genetic Algorithms

Xie Tao Zhang Yulin
(Department of Aerospace Technology, NUDT, Changsha 410073)

Abstract Much inportance has been attachad to the selection of optimal feature parameters subset in the fault diagnosis fields such as liquid rocket propulsion system. This paper presents an effective method of selection for the optimal feature parameters subset using Genetic Algorithms and based on the maximum and minimum clustering criterion for samples, so that the selected feature parameters subset can be used to compose a simplified real-time fault classifier with high robustness to various sorts of noises and disturbances. First, a composite directional divergence index for samples is proposed as an evaluation criterion for the selected feature parameters subset for fault diagnosis purpose; then, Genetic Algorithm has been modified in parts for this specific permutation problem, the dynamic fitness adaptation technique and all-sharing function are introduced in order to avoid the population's premature convergence. An ad-hoc genetic operator is specially designed to improve the feature selection efficiency. In an addition, all the selection procedures for the optimal feature parameters subset are based on the data set for 16 sorts of common faults simulated for a type of liquid rocket engine system. The numerical experiments show that this selection algorithm is highly effective and the constructed fault classifier with the selected feature parameters possesses more robustness.

Key words statistic clustering, directional divergence index for samples, feature selection, fault simulation, optimization, genetic Algorithms

液体火箭发动机的故障诊断可视为一个模式识别过程[1], 模式识别的任务是利用从样本中提取出的特征将样本划分成相应的模式类别. 研究表明[2]特征过多或过少都会降低对样本的分辨率. 因此, 如何从数以百计的发动机参数中选择尽量少的监控参数实现对常见故障的有效分离, 一直是液体火箭发动机故障诊断界关注的基本问题之一.

* 国家自然科学基金资助项目
1997年 4月 1日收稿
第一作者: 谢涛, 男, 1966年生, 博士生

本文从提高故障诊断中故障模式识别对各种噪声及干扰的鲁棒性出发,先确定故障特征参数子集的优选准则,即使类内离散度最小而类间离散度最大。然后,利用随机启发式搜索技术从所有特征参数中优选出一组参数,使得由所选故障特征参数组构造的故障模式分类器,比全特征参数故障分类器具有对故障模式更高的可分离性,并且对噪声和干扰具有更强的鲁棒性,以实现故障分离中的特征维数压缩。

参数优选是一个组合优化问题。模拟进化仿生类过程算法中的遗传算法以其强大的隐并行模式空间搜索能力[3],与随机、自适应、稳健的解空间搜索特点[9],已获得越来越广泛的应用,曾用来解决著名的"旅行商"问题[6]以及多目标函数的优化问题[4][5],且正处"升温"之中。本文利用遗传算法解决液体火箭发动机故障诊断中监控参数选择的组合优化问题。为此,我们对传统的遗传算法进行了改进,并设计了专门的遗传算子,进一步提高了其搜索效率。

1 故障特征参数优选准则

高维欧氏空间上的数据矢量几何分布的离散程度可以用欧氏距离或矢量方向的离散程度来描述。鉴于监控数据矢量方向在故障检测中的实用性,本文仍使用数据矢量方向的几何分布离散度。

高维采样数据矢量 $\boldsymbol{X}, \boldsymbol{Y} \in R^d$ 之间的方向相似度定义为它们之间夹角的余弦,即:

$$S(\boldsymbol{X}, \boldsymbol{Y}) \overset{def}{=} \cos(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\boldsymbol{X}^T \cdot \boldsymbol{Y}}{\|\boldsymbol{X}\| \times \|\boldsymbol{Y}\|}; \tag{1}$$

可见,$S(\boldsymbol{X}, \boldsymbol{Y})$是相应 $\boldsymbol{X}, \boldsymbol{Y}$ 的两个单位矢量之间的点积[10]。

虽然文[10]关于聚类结果优劣度的评估准则是目前较好的评估准则,但对于特征维数过高的样本集,总类内样本散射矩阵 $\boldsymbol{S}_W$ 不易求逆。为了避免超高维矩阵求逆,我们采用了较简单的总类内离散度与类间离散度来代替总类内散射矩阵及类间散射矩阵。

定义 $J_I = \frac{2}{n_i(n_i-1)} \sum_{\boldsymbol{X}_j \in C_i} \sum_{\boldsymbol{Y}_h \in C_i, \boldsymbol{X}_j \neq \boldsymbol{Y}_h} \cos(\boldsymbol{X}_j, \boldsymbol{Y}_h)$为类内矢量方向平均离散度,其中 C_i 表示第 i 个聚合组,$\boldsymbol{X}_j$ 与 $\boldsymbol{Y}_h$ 是属于聚合组 C_i 中的 d 维数据矢量样本,n_i 为聚合组 C_i 中样本的个数。

定义 $J_B = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \cos(\boldsymbol{C}_i, \boldsymbol{C}_j)$为类内矢量方向平均离散度,其中 C_i, C_j 表示第 i 个与第 j 个聚合组的 d 维均值样板矢量,n 为总的聚合组数。

这样,聚类结果的评估准则可简化为:

$$\max_{O \in R^d} J(O) = \frac{1}{n} \sum_{i=1}^{n} J_I - \gamma J_B \tag{2}$$

其中 $O = \{C_1, C_2, \cdots, C_n\}$为论域中样本集的 n 个 d 维聚类中心矢量,常数 $\gamma \in [0.5, 1.0]$。

对于故障特征参数的优选问题,就是从样本的特征参数集中选出一组尽量小的特征参数,使得由该组参数构成的特征矢量对原样本数据矢量集具有最大的可分性,即

$$\max_{f_{sub} \in F} J(f_{sub}) = \frac{1}{n} \sum_{i=1}^{n} J_I(f_{sub}) - \gamma J_B(f_{sub}) \tag{3}$$

其中 $f_{sub} \in F$,F 表示从总特征参数集中进行特征参数子集优选的可选总集。

假定故障诊断论域中的样本可由 N 个的特征参数来描述,即 $f = (f_1, f_2 \cdots, f_N)$,其中 f_i 为样本的第 i 个取值,可以是离散或连续的特征参数。设从 f 中优选特征参数的可能选择方案集合为 F,则 $|F| = \binom{N}{1} + \binom{N}{2} + \binom{N}{3} + \cdots + \binom{N}{N}$。

因此,特征参数子集的优选问题即为:在可选集 F 中优选出某一特征参数组 $f_{sub} \in F$,使(3)式取最大值。

2 遗传算法用于故障分离中的特征参数选择

遗传算法具有三个集合概念:种群总体(Population)、父本(Parent)、子代(Offspring)。遗传算法借

鉴生物种群发展过程中的三种进化机制: 1)物种选择 (Mate selection); 2)杂交 (Crossover or Recombination); 3)变异 (Mutation)。这三种进化机制在遗传算法中称为三个基本算子, 其作用域是上面所定义的三个集合以及集合中的个体。为提高故障特征参数选择效率, 引进一种特殊基因算子: 染色体上的基因迁移算子 (Transition Operator)。

存在两种基本种群更新机制: "父子混合选择"与"自然选择"。本文中我们采用"父子混合选择"更新机制, 即父本与子代均参与竞争下一代父本集, 每代中最优解得以保存。

$I=f=(f_1, f_2, f_3\cdots, f_{60})$表示一个个体, 其中 $f\in\{0, 1\}$, 即 $f_i=1$表示特征参数 i被选, $f_i=0$表示特征参数 i未被选。

$J(\cdot): f\longrightarrow R$ 为解串的适应值评估函数。在本文对故障特征参数的选择问题中, $J(\cdot)$即指第一节中的类内类间样本综合离散度指标, 但对于本文的优选问题, 特征参数组的大小也是一个待优化的问题, 因此本文采用如下修正的解串的适应值评估函数 $J^*(\cdot)$:

$$J^*(I)=\begin{cases}-\text{erro}, & \text{当用所选特征参数集对故障样本识别时存在 erro 个误分样本}\\ N_F\cdot J(I)/\|I\|; & \text{当 } J(I) \text{ 为正时}\\ \|I\|\cdot J(I)/N_F; & \text{当 } J(I) \text{ 为负时}\end{cases}$$

也可以选取:

$$J^*(I)=\begin{cases}N_F\cdot J(I)/\|I\|\cdot(\text{erro}+1) & \text{当 } J(I) \text{ 为正时}\\ \|I\|\cdot J(I)\cdot(\text{erro}+1)/N_F; & \text{当 } J(I) \text{ 为负时}\end{cases}$$

其中 N_F 为初始预设特征参数组的大小。

因为较小的特征参数组具有较高的适应值, 因此 $J^*(\cdot)$与所选故障特征参数组的优劣度具有单调一致的对应关系。

初始种群总体中的父本集按初始预设特征参数组的大小均匀随机地从原始解空间 $F=\{f_{sub}\}$中选取, 即使 U 中的初始个体 I_i 中的 N_F 个随机基因座上的基因为 1, 而其余其因座上的基因均置 0。

用于本文中进行故障特征参数选择的组合优化问题中各基因算子具体解释如下:

杂交算子 $R(I^*, I^{**})=I$: 对依概率随机选取的两条染色体 I^* 与 I^{**}, 相应基因座上的相同等位基因保持不变, 互异等位基因则按概率进行随机互补杂交 (即部分互补交差算子); 应用杂交算子的作用在于: 它使杂交产生的新个体不同于用作父本的个体, 使算法所产生的种群可遍历定长编码格式的解搜索空间, 以达到全局优化的目的。

基因迁移算子 $T(I)=I^*$: 对经杂交算子作用产生的染色体 I, 随机选取染色体上具有互易基因的两个基因座, 并对该二基因座上的互异等位基因进行更换 (基因交换), 且每条染色体上的迁移基因数量可选; 迁移算子的作用相当于对已有故障特征集进行优化, 即用较优的未选故障特征取代已选故障特征集中的较劣者。

变异算子 $M(I^*)\rightarrow I^{**}$: 对经迁移算子作用产生的染色体 I^*, 在父本染色体中具有相同等位基因的基因座 (在种群进化过程中相对稳定的基因块) 上, 依概率随机选取某些基因, 并使其产生随机突变 (求反操作); 就故障特征参数的选择而言, 变异算子相当于对故障特征参数个数的优化以及对已有故障特征集的优化, 即可从所有故障特征中优选出一组最小的特征参数组。

杂交父本选择采用轮盘赌策略, 以与父本集合中个体 I_K 的性能指标评估值 $J^*(\cdot)$成比例的概率 P_x, 随机选取父本集合中的个体 I_X, 其中 $P_x=J^*(I_x)/\sum_{i=1}J^*(I_i)$。

为了防止遗传算法的过早收敛, 还可以采取动态改变每代中父本的适应值并采取"群集"技术或引进共享函数, 以限制父本中的所有个体趋于一致[7]。本文中将每代中的父本的适应值减去其最小值, 动态地得到每一父本与父本中适应值的最小值的差值作为该父本的动态适应值, 再利用共享函数对所有父本的动态适应值进行修改。共享函数对 I_i 适应值按下式修改: $J^{**}(I_i)=J^*(I_i)/\sum_{j=1}\text{sh}(S(I_i, I_j))$, 其中 $S(I_i, I_j)$是描述 I_i 与 I_j 相似程度的尺度函数, $\text{sh}(S(\cdot, \cdot))$是共享函数。这样, 在相邻区域内的个体

相互增加了其共享函数值, 限制了较相似的个体在群体中的增长, 从而保持了群体的多样性, 避免了一旦发现最优点, 遗传算法就引导群体聚集到它附近, 而将其它信息逐渐抛弃。

此外, 改进的遗传算法中所使用的杂交概率、基因迁移概率以及变异概率都是自适应的[8]。

3 数值实验结果

分析表明, 某型号液体火箭发动机主级段工况下的常见故障为 16大类。通过对该型号发动机建立的主级段工况下的数学模型, 我们仿真了 16大类常见故障导致 60个特征参数出现故障过渡过程的数据集, 其中对每类故障仿真了 40组故障等级渐增的数据集 (假定故障都为缓变型)。若以每组数据的中心矢量方向为相应故障类的样板矢量方向, 则数值分析表明整个仿真数据集 (640个 60维的数据)是 100% 可分的。为了简化故障分离的计算复杂度以及提高故障模式类的可分性, 需从 60个发动机参数中优选出一组尽量小的参数集, 且不降低其对已知故障模式的可分性。为提高样本的可分离性, 所有样本数据都经过故障偏差的标准化处理: 设 P_0 为某特征参数的主级段正常工况值, P 为该参数的故障仿真值 (或故障状态下的测量值), 那么 $\overline{P}=\dfrac{P-P_0}{P_0}$; $\overline{P}$ 称为该参数的故障偏差标准化值。

我们预设初始特征参数集大小 N_F 为 60, 并取父本集合大小为 6, 每个父本平均赋予 10个子代个体繁植机会。图 1表示所选故障特征参数子集随遗传算法种群进化代数发展的适应性调整情况, 图中曲线旁边的数据指最终所选特征参数的序号。图 2表示用所选特征参数子集对故障仿真样本数据集进行故障模式分离时, 其误分个数随遗传算法种群进化代数发展的降低情况。图 3表示由所选故障特征参数子集构成故障分类器时, 故障类均值模板矢量方向之间的平均夹角大小随遗传算法种群进化代数发展

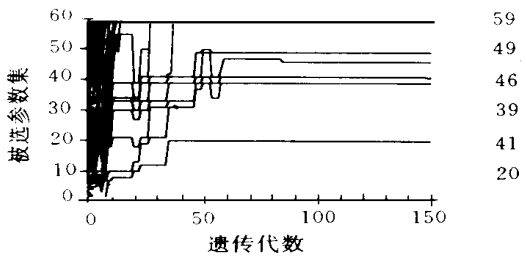

图 1 被选特征参数随遗传算法代数发展情况

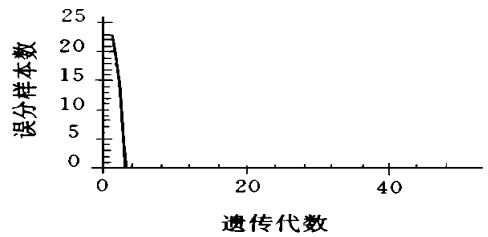

图 2 误分样本数随遗传代数发展情况

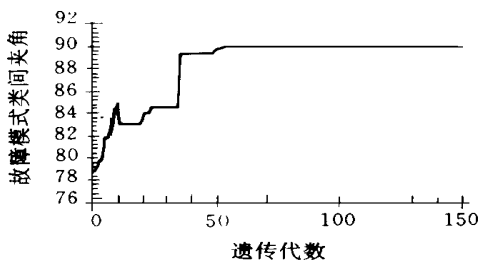

图 3 故障模式类间夹角 (由相似度换算得到)随遗传代数发展情况

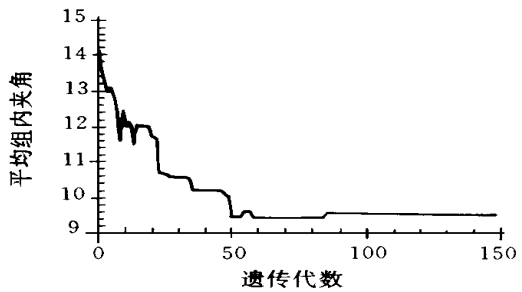

图 4 平均组内样本夹角 (由相似度换算得到)随遗传算法代数发展情况

的变化情况, 可见各故障类间的可分离性随种群的进化而提高。图 4表示由所选故障特征参数子集构成故障分类器时, 故障类内样本矢量方向之间的总平均夹角大小随遗传算法种群进化代数发展的变化情况, 可见各故障类内样本矢量方向的一致性随种群进化而提高。图 5表示由所选故障特征参数子集构成故障分类器时, 类内类间样本的综合离散度与所选特征参数子集大小的关系。图 3 图 4与图 5表明, 故障样本中故障类的可分离性随特征参数集的优化而提高。由于种群更新过程的随机性, 以上各曲线均为 5次实验的平均结果。

故障特征参数集的选择过程是, 先使由所选特征参数子集构造的分类器达到全部样本数据可分, 然后再在该可分特征参数子集的基础上优化其可分性能, 提高按所选特征参数子集而建立的故障分类器对各种噪声干扰的鲁棒性.

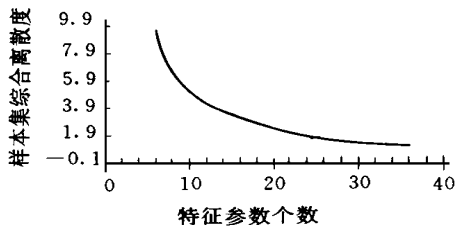

图 5 故障样本集的综合离散度与特征参数个数之关系

最后, 由所选故障特征参数组构造故障分类器, 对由同一故障仿真程序仿真出来的 16类故障等级不同的 1000个模拟故障数据样本进行故障模式分离实验, 结果无误分样本出现, 因此可认为对样本测试集的识别率达 100%.

4 结论与讨论

数值实验结果表明, 该特征参数选择算法具有理想的参数选择效果, 改进的遗传算法对于本文的组合优化问题具有极高的解空间搜索效率, 同时该优选准则能使由所选特征参数构成的故障分类器对故障样本集具有最大的可分离度. 此外, 数值实验结果还表明, 故障的可分离性与所选特征参数子集的大小不成简单的正比关系.

本文采取了动态调整父本的适应值并引进共享函数, 且解的染色体编码中每个基因座上仅具有二个等位基因, 基本上避免了局部最优组合死区的出现.

类内类间样本综合离散度不是唯一的优选准则. 基于统计意义上的 Shannon互信息定义, 可以较好地测度特征参数所含的故障类别信息以及特征参数与特征参数之间的信息冗余度, 可以使优选特征参数组中特征参数所含的平均故障分类信息尽量多, 而特征参数与特征参数之间的平均信息冗余度尽量小, 以此作为优选原则. 基于 Shannon互信息定义的特征参数优选准则, 可以很好地完成对故障特征参数子集的优选问题, 而且表现出与类内类间样本综合离散度特征参数子集优选准则较好的一致性[11].

参考文献

1 Nemeth E And Norman A M. Development Of A Health Monitoring Algorithm. AIAA/SAE/ASME/ASEE 26th Joint Propulsion Conference, July 16~18, 1990/Orlando, FL

2 Tou J and Gonzalez R. Pattern Recognition Principles. Addison Wesley, Reading, Massachusetts 1974

3 Alberto Bertroni, Marco Dorigo. Implicit parallelism in genetic algorithms. Artificial Intelligence 61, 1993: 307~314

4 Hajela P, Lin C Y. Genetic search strategies in multicriterion optimal design. Structural Optimization, 1992, 5(4): 99~107

5 Schaffer J D. Multiple objective optimization with vector evaluated genetic algorithms. In: Proc. of International Conference on Genetic Algorithms and Their Applications. 1985: 93~100

6 Ulder N L J, Aartselt E H L. Genetic local search algorithms for the traveling salesman problem. In: paralel Problem Solving From Natrue H.-P. Schwefel-R. Manner Eds., Springer-Verlag, 1990: 109~116

7 Goldberg D E, Richardson J. Genetic algorithms with sharing for the multimodel function optimization. In: Proc. of the 2nd Interoni Confeence

8 Srinivas M, Patnail L M. Adapative probabilltis of crosover and mutation in genetic algorithms. Trana. Sys. Man and Cybern. 1994, 24(4): 645~667

9 徐宗本, 李国. 解全局优化问题的仿生类算法(I)-模拟进化算法. 运筹学杂志, 1995, 16(2)

10 李金宗. 模式识别导论. 高等教育出版社, 1994