

演化策略用于高维故障样本集最优统计聚类分析*

谢涛¹, 陈火旺¹, 张育林²

(1. 国防科技大学 计算机学院, 湖南 长沙 410073;
2. 国防科技大学 航天与材料工程学院, 湖南 长沙 410073)

摘要: 针对液体火箭发动机推进系统超高维故障样本数据的聚类问题, 提出基于演化策略的最优统计聚类算法。为预防算法过早收敛, 演化策略采用了父本适应值的动态调整值与共享函数, 并针对超高维数据聚类提出了控制参数的适应性调整技术; 为使算法能最终跳出局部最优死区, 提出算法的局部调整策略。该算法用于液体火箭发动机典型故障仿真数据集分析, 并取得了最优聚类结果。此外, 还基于 IRIS 数据集比较了该算法与 FKCN 模糊自主聚类算法。仿真分析表明了算法在高维数据聚类分析中的优点。

关键词: 液体推进剂火箭发动机; 故障诊断; 故障分析; 演化策略; 计算数学

中图分类号: O241; V434.1 文献标识码: A 文章编号: 1001-4055(2000)05-0034-05

Optimal statistical clustering for high dimensional fault sample using evolution strategies

XIE Tao¹, CHEN Huo-wang¹, ZHANG Yu-lin²

(1. Inst. of Computer Science, National Univ. of Defense Technology, Changsha 410073, China;
2. Inst. of Aerospace & Material Engineering, National Univ. of Defence Technology, Changsha 410073, China)

Abstract: A clustering algorithm based on Evolution Strategies was proposed to make analysis on high dimensional data of liquid rocket engine propulsion system. In order to prevent the solution population from premature convergence, the dynamic fitness adaptation technique and all-sharing function were introduced. An adaptive regulation scheme for evolution control parameters was specially presented for clustering analysis of high dimensional data. A local clustering deadlock can also be overcome by the deadlock check and cluster recombination & collapse strategies. This algorithm was used in the optimal clustering analysis for the 560 data samples of 14 sorts of common faults, each of 68 dimensions, which were simulated for a liquid rocket engine. In addition, comparison with fuzzy Kohonen clustering networks (FKCN) has also been made, based on IRIS data. The simulation results show that the evolution strategies based on clustering algorithm is superior to other non evolutionary clustering algorithms, particularly when the data is of high dimensions.

Key words: Liquid propellant rocket engine; Fault diagnosis; Fault analysis; Evolution strategies; Computational mathematics

1 引言

科学研究与工程实践中经常遇到超高维数据的最优统计聚类问题, 而常规动态聚类算法诸如 ISO-DATA 算法、模糊自主聚类算法 (FKCN) 等等, 都不能有效解决这一类问题。遗传算法虽然可用于解决这类问题, 但对于较高聚类组数问题, 聚类效率

偏低^[1]。演化策略^[2~4]借鉴生物自然选择与遗传变异机制, 是一种随机、自适应、稳健的搜索算法, 特别适用于大规模、非线性、多峰、甚至无目标函数表达式的优化问题, 演化操作算子直接作用于实变量本身, 算法在问题可行解空间中直接搜索, 已成功应用于超高维函数以及多目标函数的优化问题。本文在聚类算法和演化算法的基础上^[5,6], 利用演

* 收稿日期: 1999-10-23; 修订日期: 2000-01-17。基金项目: 国家自然科学基金资助项目 (NSF69785002)。

作者简介: 谢涛 (1966—), 男, 博士后, 副教授, 研究领域为大型液体火箭发动机的故障检测与诊断技术。

化策略对液体火箭推进系统高维故障样本数据集进行最优分区聚类。

2 系统典型故障数据集的聚类问题

聚类算法在文献 [5] 中已有介绍。本文聚类对象是液体火箭推进系统的测量数据, 设

$V = \{v_1, v_2, \dots, v_n\}$ 其中 v_i 为 d 维采样数据矢量, $1 \leq i \leq n$;

$O = \{c_1, c_2, \dots, c_m\}$, 其中 c_i 为 d 维聚类中心数据矢量, $1 \leq i \leq m$ 。

故障诊断中样本相似度一般采用数据矢量方向相似度。高维采样数据矢量 $x, y \in R^d$ 之间的方向相似度定义为它们之间夹角的余弦^[5], 即

$$S(x, y) \triangleq \cos(x, y) = \frac{x^T \cdot y}{\|x\| \cdot \|y\|} \quad (1)$$

$S(x, y)$ 也是相应 x, y 的两个单位矢量之间的点积。可以证明^[1], 方向相似度与数据矢量中元素值成单调升关系, 因此聚类过程中应标准化各测量参数, 使得矢量方向相似度对温度、转速、流量、压力等不同参数具有基本相同的敏感度。

基于演化策略的聚类算法可采用硬聚类评估函数或模糊聚类评估函数作为解的适应值函数。基于矢量方向相似度定义的聚类目标函数^[5], 可取

$$\max J(W, O) = \sum_{i=1}^m \sum_{j=1}^n u_{ij} \cos(v_j, c_i) \quad (2)$$

其中 $W = (u_{ij})_{n \times m}$ 隶属矩阵, 且 $\sum_{i=1}^m u_{ij} = 1$ 。当 $u_{ij} \in \{0, 1\}$ 时, 相应聚类算法为硬聚类算法; 当 $u_{ij} \in [0, 1]$ 时, 相应聚类算法为模糊聚类算法。 $L(l) \leq c_i(l) \leq R(l)$, $1 \leq l \leq d$, $1 \leq i \leq m$; $c_i(l)$ 为聚类中心数据矢量 c_i 的第 l 个元素, $L(l) = \min_j (v_j(l))$, $R(l) = \max_j (v_j(l))$, $1 \leq j \leq n$ 。

迭代终止条件取 $\sum_{i=1}^m \cos(c_{i,t}, c_{i,t-1}) \geq m - \varepsilon$ 或 $|J(W_t, O_t) - J(W_{t-1}, O_{t-1})| \leq \varepsilon$ 其中 ε 取较小的正数 (如取 $\varepsilon = 1 \times 10^{-6}$)。

设 W^* 和 O^* 是相应全局最优聚类结果, 则 W^* 和 O^* 可以相互转化。有两种解决该聚类问题的途径: (1) 搜索全局最优解 W^* 来优化 $J(\cdot, \cdot)$; (2) 搜索全局最优解 O^* 来优化 $J(\cdot, \cdot)$ 。对于硬聚类算法, 前者属于组合优化问题, 可用遗传算法解决^[1]; 后者属于连续量优化问题, 可以用演化策略、模拟退火算法、FKCN 等算法来解决^[1,5,6]。

设实数串 $X^T = (c_1^T, c_2^T, c_m^T)$ 表示基于演化策

略的聚类算法的一个解, 即由 m 个聚类中心数据矢量按序排列。适应值评估函数 $f(X)$ 可如下计算:

(1) 通过数据集 V 与聚类中心矢量 c_i ($1 \leq i \leq m$) (由 X^T 得到), 得到隶属矩阵 $W = (u_{ij})$ 。

对于硬聚类算法

$$u_{ij} = \begin{cases} 1; & \text{如果 } \cos(v_j, c_i) = \cos\{\max_i (v_j, c_i)\} \\ 1 \leq i \leq m, 1 \leq j \leq n. \\ 0; & \text{其它。} \end{cases}$$

对于模糊聚类算法^[1,5]

$$u_{ij} = \frac{1 + \cos(v_j, c_i)}{\sum_{k=1}^m (1 + \cos(v_j, c_k))}$$

(2) 根据新的隶属矩阵 W 更新聚类中心矢量 c_i , $1 \leq i \leq m$ 。

$$c_i = \frac{\sum_{j=1}^n u_{ij} v_j}{\sum_{j=1}^n u_{ij}}$$

(3) 根据式 (2) 求出 $f(X) = J(W, O)$, 即得解 X^T 的适应值。

3 高维数据聚类问题的演化策略

Schwefel^[3,4] 系统推广了 Rechenberg 的原始演化策略^[3], 建立了 $(\mu + \lambda) - ES$ 及 $(\mu, \lambda) - ES$ 演化策略。

定义父本集合 $U = \{I_1, I_2, I_\mu\}$, 子代集合 Θ , 种群总体集合 $\Pi = U \cup \Theta$, 并且设 $|U| = \mu$, $|\Theta| = \lambda$ 其中 $|U|$ 和 $|\Theta|$ 分别表示集合 U 与集合 Θ 中元素的个数。 $I = (X, \Delta)$ 表示解个体, 其中 $X^T = (c_1^T, c_2^T, \dots, c_m^T)_{m \times d}$, $\Delta^T = (\sigma_1^T, \sigma_2^T, \dots, \sigma_m^T)_{m \times d}$ 为矩阵解 X^T 的变异方差矩阵, 其元素的初始值与相应测量参数的噪声方差成正比, 标准化后取 $\sigma_j = 0.01$, 其中 $1 \leq i \leq m$, $1 \leq j \leq d$ 。

$f(\cdot): X_{m \times d}^T \rightarrow R$ 为解串适应值评估函数, 其计算过程如二节所示。

中和重组算子 $R(I^*, I^{\ddot{}}) \rightarrow I^{\dot{}}$: 设 $I^* = (X^*, \Delta^*)$, $I^{\ddot{}} = (X^{\ddot{}}, \Delta^{\ddot{}})$, $I^{\dot{}} = (X^{\dot{}}, \Delta^{\dot{}})$; 在 m 个聚类中心矢量中均匀随机选取 k_c 个中心矢量 c_i , 再从每一 c_i 中随机均匀选取 l_c 个元素 c_{ij} , 其中 $1 \leq k_c \leq m$, $1 \leq l_c \leq d$; 使得 $c_{ij} = \frac{1}{2}(c_{ij}^* + c_{ij}^{\ddot{}})$, $\sigma_{ij} = \frac{1}{2}(\sigma_{ij}^* + \sigma_{ij}^{\ddot{}})$, $1 \leq i \leq m$, $1 \leq j \leq d$ 。变异算子 $M(I^{\dot{}}) \rightarrow I^{\dot{}}$: 设 $I^{\dot{}} = (X^{\dot{}}, \Delta^{\dot{}})$; 在 m 个聚类中心矢量中均匀随机选取 k_m 个中心矢量 c_i , 再从每一 c_i 中随机均匀选取 l_m 个元素 c_{ij} , 其中 $1 \leq k_m \leq m$, $1 \leq l_m \leq d$; 使得: $\sigma_j =$

$\sigma_{ij} \exp(N_o(0, v))$, $c_{ij} = \dot{c}_{ij} + N_o(0, \alpha_{ij})$, $\forall i, j, 1 \leq i \leq m, 1 \leq j \leq d$. 其中 $N_o(0, v)$ 是方差为 v 的高斯函数, v 为全局变异方差步长, 随种群代数逐渐减小。

为了提高高维数据的聚类效率, 加快收敛速度, 必须对重组与变异操作的控制参数 k_c, l_c, k_m, l_m 分别作适应性调整, 即使得重组与变异算子的作用范围随每代中父本平均适应值的变化, 逐步收敛至最佳聚类中心矢量的有效聚类元素上。设 $I_i(k)$ 为第 k 代父本中个体 i , $k_c(k), l_c(k), k_m(k), l_m(k)$ 分别为第 k 代中重组与变异操作的控制参数, 初值分别取 $k_c(0) = k_m(0) = m, l_c(0) = l_m(0) = d$. 控制参数随进化过程的调整策略可用伪程序语言描述如下:

IF $(\frac{1}{\mu} \sum_{i=1}^{\mu} f(I_i(k+1))) > \frac{1}{\mu} \sum_{i=1}^{\mu} f(I_i(k))$.

THEN $\{k_c(k+1) = k_c(k), l_c(k+1) = l_c(k),$

$k_m(k+1) = k_m(k), l_m(k+1) = l_m(k)\}$.

ELSE $\{k_c(k+1) = k_c(k), l_c(k+1) = l_c(k) - 5,$

$k_m(k+1) = k_m(k), l_m(k+1) = l_m(k) - 5\}$.

IF $(l_c(k+1) \leq 5$ 或 $l_m(k+1) \leq 5,$

且 $k_c(k+1) \geq 2$ 与 $k_m(k+1) \geq 2$).

THEN $\{k_c(k+1) = k_c(k) - 1, k_m(k+1) =$

$k_m(k) - 1, l_c(k+1) = l_c(0), l_m(k+1) = l_m(0)\}$.

最优统计聚类分析的演化策略基本步骤为:

(1) 随机产生个初始父本个体 $I_i(0) = (X_i, \Delta)$, $1 \leq i \leq \mu$; 其中 X_i 由从故障样本集中随机抽取的 m 个数据矢量按序排列构成, 计算其适应值并按序排列;

(2) 基于父本被选概率 $P_x = f^*(I_x(k)) / \sum_{i=1}^{\mu} f^*(I_i(k))$, 按轮盘赌规则随机选取 λ ($\lambda > \mu$) 对父本以重组与变异控制参数 $k_c(k), l_c(k), k_m(k), l_m(k)$ 进行中和重组与变异, 产生 λ 个子代个体, 计算其适应值并按序排列;

(3) 按 $(\mu + \lambda) - ES$ 更新机制选取下一代父本;

(4) 按控制参数调整策略调整下一代重组与变异控制参数 $k_c(k+1), l_c(k+1), k_m(k+1)$ 与 $l_m(k+1)$;

(5) 判断最优父本个体是否满足问题要求或已完成指定演化代数? 是, 则终止种群更新循环, 输出其结果; 否则, 转步骤 (2), 继续种群更新循环。

为了防止算法过早收敛, 本文采用父本动态适

应值与共享函数。共享函数对 I_i 适应值按下式修改: $f^*(I_i) = f(I_i) / \sum_{j=1}^{\mu} SM(I_i, I_j)$; 其中 $SM(I_i, I_j) = SM(x_i, x_j) = \frac{1}{m} \sum_{k=1}^m S(c_{i,k}, c_{j,k})$ 。采用共享函数可以保持群体的多样性。

4 仿真故障数据集的聚类分析

分析表明某液体火箭发动机主级段常见故障为 14 大类。通过对该发动机建立主级段的故障数学模型, 仿真了 14 大类常见故障导致 68 个监控参数出现故障过渡过程的数据集, 其中每类故障仿真了 40 组故障等级渐增的数据集, 所有数据经过偏差标准化处理。如以每组数据的均值矢量方向为相应故障类的模板矢量方向, 数值分析表明整个仿真数据集 (560 个 68 维的数据) 是 100% 可分的。为了符合实际情况下故障模式类模板矢量方向的统计过程, 并检验演化策略的聚类效率, μ 个初始父本个体中的 X_i ($1 \leq i \leq \mu$) 分别由故障样本集中随机抽取的 m 个数据矢量构成, 考察样本误分个数与演化代数的关系。

图 1 表示仿真故障数据集中误分样本数 n 随演化代数 g 的变化关系, 图 2 表示聚类目标函数值 $J(W, O)$ 随演化代数 g 的变化关系。从图 1 与图 2 可看出在最初几个进化过程中目标函数值急剧上升, 误分个数显著降低, 以后随演化代数的发展进化过程相对变缓。这说明样本集中分辨性较好的样本比分辨性较差的样本能获得较快聚类。从统计观点上说, 随着演化代数的发展, 样本集样本聚类速度与样本集总体可分辨性成比例, 而最后未获正确聚类的样本则属于聚类组之间的临界点集。

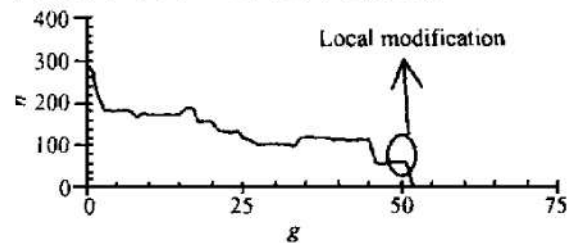


Fig. 1 Number of misclassified fault samples vs generations

用于超高维样本集聚类分析的演化策略可能因样本可分辨率不一致而陷入某一局部最优分区。如果聚类过程中存在有组内样本相似度小于或等于另外两组或两组以上混合集的样本相似度, 即可能产生局

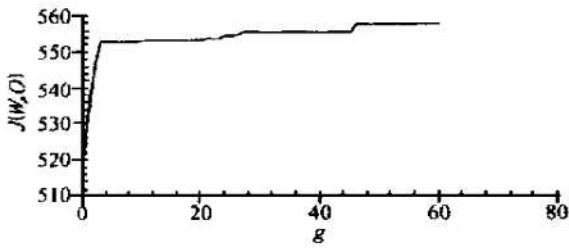


Fig. 2 Clustering objective function $J(W, O)$ vs generations

部最优分区。演化策略主要靠变异算子来探索新解，对于超高维数据聚类问题，要打破这种僵局而发现新的较好解，必须在相应解的较多指定位置上的随机变异满足要求，而这种随机组合概率是很小的，因此很难或基本上不可能仅靠随机变异算子去打破这种僵局。本文采取局部调整策略，即如果演化策略中的父本集趋于一致，且接连几代演化过程中均未发现更好解，即可确定演化策略已停滞于局部最优聚类死区。如果发现局部最优聚类死区中存在有两较小组的组间样本相似度大于某一较大组的组内样本相似度，就合并该两组较小组，即任取其中一组的中心矢量为该两组中心矢量的平均值，并设另一组的中心矢量为较大组的中心矢量加上微小的高斯随机扰动，这就是所谓的小组合并、大组分开策略。

使用局部调整策略有时会使目标函数值倒退，但这是值得付出的代价。数值实验结果表明，上述小组合并、大组分开策略能使演化策略有效地避开局部最优死区。当演化策略算法进行约 40 代后，算法已基本陷入局部最优聚类死区，采用局部调整策略后，演化策略又从第 46 代开始继续前进；同样，在第 51 代通过局部调整后，样本集所有样本均获正确聚类，虽然目标函数值稍有减少。

IRIS 数据是用来检验统计聚类算法优劣程度的一组标准数据^[7,8]。图 3 和图 4 是分别用演化策略与 FKCN 聚类算法对 IRIS 数据进行聚类的结果，其中样本相似度均采用矢量方向夹角的余弦值度量。大

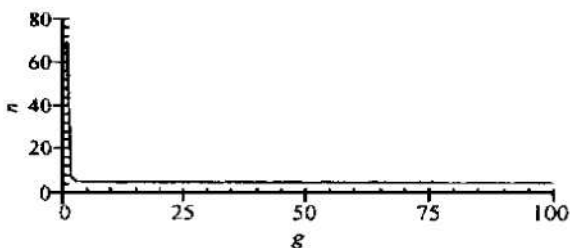


Fig. 3 Number of misclassified IRIS samples vs generations (Cosine)

量仿真结果表明，基于演化策略的聚类算法的收敛速度与聚类精度均高于 FKCN 算法，且 FKCN 的聚类过程出现振荡现象，这说明 FKCN 具有“过渡学习”的缺点。由于 IRIS 数据是一组低维^[4]数据，因此不必对演化控制参数作适应性调整。

此外，仿真结果还表明，对于类似 IRIS 数据的低维数据样本，如果样本间相似度采用欧氏距离度量，基于演化策略的聚类算法收敛速度低于 FKCN 聚类算法，但聚类结果稍优于 FKCN 算法，如图 5 与图 6 所示。

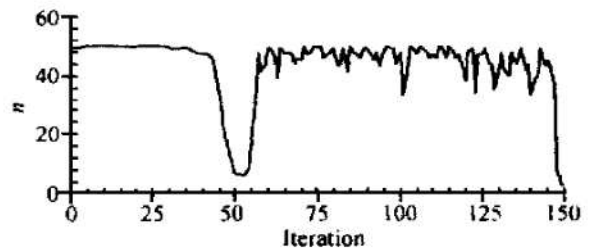


Fig. 4 Number of misclassified IRIS samples vs FKCN iterations (Cosine)

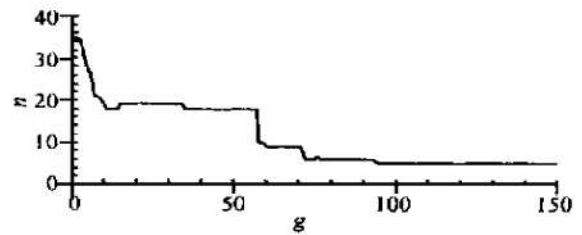


Fig. 5 Number of misclassified IRIS samples vs generations (Euclidean)

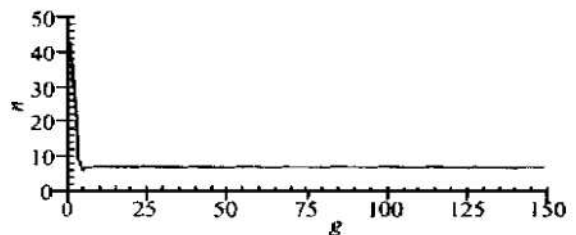


Fig. 6 Number of misclassified IRIS samples vs FKCN iterations (Euclidean)

5 结 论

对 560 个 14 类 68 维的超高维故障仿真数据样本，进行了演化策略最优统计分区聚类分析，并基于 IRIS 数据的聚类结果与 FKCN 进行了比较。仿真分析结果表明，基于演化策略的最优统计聚类算法对高维数据集的聚类分析效率高于传统非进化类聚类算法，但在低维数据（如 IRIS 数据）聚类分析上二者并无明显差别。
(下转第 52 页)

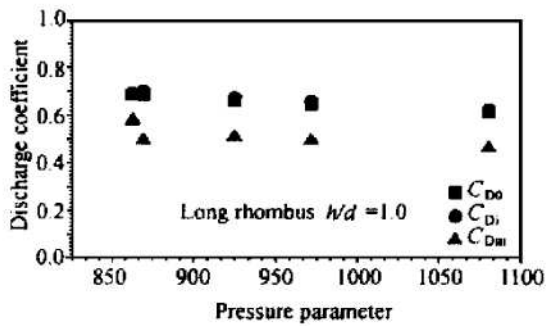


Fig. 7 No. 1 model

此时压力参数对流量系数基本无影响。双层壁中，冲击壁流量系数大于多斜孔壁流量系数。原因是多斜孔的孔长径比和倾斜进气损失比垂直进气的冲击壁大。

(4) 对不同几何结构实验板的对比研究表明，双层壁实验板长菱形排布的流量系数值高一些。双层壁夹缝高度的改变对流量系数的影响很小。

主燃区冲击加多斜孔双层壁设计时，因压力参数小，流量系数值较低，对开孔孔径有较大影响，应避免因开孔面积而导致主燃区冷却气流量较小。设计孔径时需要仔细考虑。

参考文献

- [1] 高潮, 褚孝荣, 王宝官. 燃烧室壁冲击冷却换热的实验研究 [J]. 推进技术, 1998, 19 (1).
- [2] 林宇震, 李彬, 宋波, 等. 多斜孔壁冷却方式不同进气角度小孔内对流换热研究 [J]. 推进技术, 1999, 20 (1).
- [3] Andrews G E, Mkpadi M C. Full coverage discrete hole wall cooling: discharge coefficients [R]. ASME 83-GT-79.
- [4] Hay N, Lampard D, Benmansour S. Effect of crossflow on the discharge coefficient of film cooling holes [R]. ASME 82-GT-147.
- [5] 高峰. 多斜孔流量系数研究及冷却技术分析 [D]. 北京: 北京航空航天大学, 1995.
- [6] 方韧. 燃烧室多斜孔壁流量系数研究 [J]. 航空动力学报, 1998 (1).
- [7] Champion J L, et al. Experimental investigation of the wall flow and cooling of combustion chambers walls [R]. AIAA 95-2498.
- [8] Tay Chu, et al. Discharge coefficients of impingement and film cooling holes [R]. ASME 85-GT-81.

(责任编辑: 盛汉泉)

(上接第 37 页)

通过采用父本的动态适应值、引进共享函数并对演化过程控制参数作适应性调整, 虽然不能完全避免局部最优聚类死区, 但能延迟聚类死区的到来。当聚类死区最终出现时, 采取聚类死区检测与小组合并、大组分裂策略使算法跳出聚类死区。数值实验也证实了这一点。

参考文献

- [1] 谢涛. 基于进化计算的液体火箭发动机故障诊断研究 [D]. 长沙: 国防科技大学, 1998.
- [2] Rechenberg I. Evolutionstrategie: optimierung technischer systeme nach prinzipien der biologischen evolution [M]. Stuttgart: Frommann-Holzboog Verlag, 1973.
- [3] Schwefel H P. Numerische optimierung von computer-modellen mittels der evolutionstrategie [M]. Interdisciplinary systems

Research, Vol. 26, Basel: Birkhuser, 1977.

- [4] Schwefel H P. Numerical optimization of computer models [M]. Chichester, UK: John Wiley, 1981.
- [5] 谢涛, 张育林. GA-HCM 混合聚类算法及其在液体火箭发动机故障检测中的应用 [J]. 推进技术, 1997, 18 (1).
- [6] 谢涛, 张育林. 非线性混合回归演化算法研究 [J]. 推进技术, 1998, 19 (5).
- [7] Eric Chen Kuo Tsoa, Bezdek J C, Nikhil R Pal. Fuzzy kohonen clustering networks [J]. Pattern Recognition, 1994, 27 (5): 757~764.
- [8] Phanendra Babu G, Narasimha Murty M. Clustering with evolution strategies [J]. Pattern Recognition, 1994, 27 (2): 321~329.

(责任编辑: 盛汉泉)