基于聚类分析和可视化的增强遗传算法
—I．算法的引出、原理与分析

王克峰， 孙晓静， 姚平经

(大连理工大学化工学院化学系统工程系，辽宁 大连 116012)

摘 要：提出了一种基于数据可视化的聚类分析法(Cluster Constrained Mapping, CCM)和人机结合的增强遗传算法，可保证进化过程在可行域中进行，不需要任何惩罚函数参数，可有效地进行带约束问题的优化.

关键词：可视化；聚类分析；带约束优化；遗传算法

中图分类号：TP391.75 文献标识码：A 文章编号：1009–606X(2004)05–0438–07

1 前 言

一个约束优化问题通常由如下形式的非线性规划(Non-Linear Programming, NLP)模型来表示：

$$\begin{aligned} &\text{Minimize} && f(X) \\ &\text{Subject to} && g_{\mathrm{q}}(X) \geq 0, \qquad q = 1,\cdots,Q \\ &&& h_{\mathrm{m}}(X) = 0, \qquad m = 1,\cdots,M. \\ &&& x_i^l \leq x_i \leq x_i^u, \qquad i = 1,\cdots,n \end{aligned} \tag{1}$$

上述 NLP 模型中 $f(X)$为目标函数，有 n 个变量，Q 个大于等于类的不等式约束和 M 个等式约束.

遗传算法(Genetic Algorithms, GA)对上述 NLP 模型的优化已取得了重大进展[1]，但 GA 处理约束的惩罚函数法需要大量的实验才能确定适当的参数，并由此定义惩罚函数. 某些惩罚函数因子定义的方法在一些问题上可能效果较好，但对另一些问题就可能不令人满意[2,3]. 现在还没有一个通用的方法，这在很大程度上限制了 GA 的应用.

本研究发展了一种改进的遗传算法(Improved Genetic Algorithms, IGA). 给定一个 NLP 模型，IGA 通过数据可视化的聚类分析来获得可行域信息；重新定义进化算子(交叉和变异)，利用已获得的可行域信息，保证子代有较多的可行解；如果个体不可行，置该目标值为无穷大(如 10^{20}).

聚类分析法[4]通常定义在 n 维(n–D)空间产生聚类信息. 然而，对于大规模问题，这些方法对计算资源要求较高，结果对参数的初始值有较强的依赖性，而且其信息的传统树形表述法对用户来说往往太复杂而很难被理解并运用到以后的优化过程中. 也有一些方法在缩减的维空间中(如 2 维)来产生聚类信息，如 Kohonen 自组织特征映射(Self-Organising Map, SOM)法[5]. 这些方法易于解释，简单，并且可视化，然而，其依赖于由数据集组成的拓扑图，而且将基于 2 维的聚类信息转化到 n 维很困难.

本研究提出了一种新的基于可视化的聚类分析法来获得可行域信息. 首先，用人工神经网络(Artificial Neural Network, ANN)定义聚类约束映射(Cluster Constrained Mapping, CCM)，将数据从 n

收稿日期：2003–09–26, 修回日期：2003–11–24
基金项目：国家 973 计划资助项目(编号：G2000263)
作者简介：王克峰(1972–)，男，辽宁省大连市人，副教授，博士，化工过程系统工程专业；孙晓静，通讯联系人，E-mail: minipaula@163.com.

维空间映射到 2 维空间，然后在 2 维空间中用凝聚算法[6]进行聚类分析，获得聚类信息. 算法参数可通过可视化由用户交互提供. 最后，再将从 2 维空间得到的聚类信息映射回 n 维空间，从而得到 NLP 模型的可行域信息，用 IGA 进行优化. 一个简单非线性约束优化例子表明了算法的有效性.

2　多维数据可视化聚类分析

本研究提出的基于降维的聚类分析算法有以下 4 个步骤：

(1) 随机产生 10^n 个节点，并选出可行点来表征 n 维数集.

(2) 应用聚类约束映射(CCM)将 n 维数据映射到 2 维，实现可视化.

(3) 应用聚类分析由可视化数据获得可行域信息.

(4) 将 2 维空间的可行域信息映射回 n 维空间.

2.1 可行点的选择

给定一个 n 维，包含 N 个可行点的问题，其计算复杂度很大程度上取决于 n 和 N. 由于数据的维数 n 已被缩减到 2，则计算量由最小化 N 决定.

2.2 应用 ANN 的 CCM

CCM 的目的是在降维过程中自动保存每一对数据点的拓扑信息，使每对数据点间距离在转换前和转换后保持相同. 在训练人工神经网络中，目标函数(即划分聚类的约束)要尽可能降低距离比率值方差(E)，即：

$$E=\frac{2}{N(N-1)}\sum_{p=2}^{N}\sum_{p'=1}^{p-1}\left(\frac{\left\|O_p-O_{p'}\right\|/2}{\left\|X_p-X_{p'}\right\|/n}-1\right)^2. \tag{2}$$

前向 ANN 有 3 层：输入层、隐藏层和输出层，如图 1 所示. 图中 x_i 为给定矢量 X 的第 i 维坐标，o_k 为输出矢量 O 的第 k 维坐标，本文中 k=2. w_{kj} 是输出层 k 和隐藏层 j 间的权值，w_{ji} 为隐藏层 j 和输入层 i 间的权值.

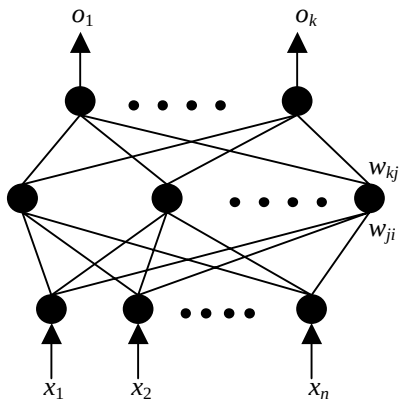

图 1　用于训练的人工神经网络(ANN)结构

Fig.1 The structure of artificial neural network (ANN) used for training

在训练 ANN 的反馈过程中，更新权值的表达式由误差 E 对权值取微分得到，如下式所示：

$$\Delta w_{kj}=-\eta\frac{\partial E}{\partial w_{kj}}=-\eta\frac{4}{N(N-1)}\frac{n^2}{2^2}\sum_{p=2}^{N}\sum_{p'=1}^{p-1}\left(\frac{\left\|O_p-O_{p'}\right\|}{\left\|X_p-X_{p'}\right\|^2}\frac{\partial}{\partial w_{kj}}\left\|O_p-O_{p'}\right\|\right)+\eta\frac{4}{N(N-1)}\frac{n}{2}\sum_{p=2}^{N}\sum_{p'=1}^{p-1}\left(\frac{1}{\left\|X_p-X_{p'}\right\|}\frac{\partial}{\partial w_{kj}}\left\|O_p-O_{p'}\right\|\right), \tag{3}$$

$$\Delta w_{ji}=-\eta\frac{\partial E}{\partial w_{ji}}=-\eta\frac{4}{N(N-1)}\frac{n^2}{2^2}\sum_{p=2}^{N}\sum_{p'=1}^{p-1}\left(\frac{\left\|O_p-O_{p'}\right\|}{\left\|X_p-X_{p'}\right\|^2}\frac{\partial}{\partial w_{ji}}\left\|O_p-O_{p'}\right\|\right)+\eta\frac{4}{N(N-1)}\frac{n}{2}\sum_{p=2}^{N}\sum_{p'=1}^{p-1}\left(\frac{1}{\left\|X_p-X_{p'}\right\|}\frac{\partial}{\partial w_{ji}}\left\|O_p-O_{p'}\right\|\right), \tag{4}$$

其中η为 ANN 的学习速率，文中给定为 0.5.

2 个输出点之间的距离对两权值的偏微分由下式给出：

$$\frac{\partial}{\partial w_{kj}}\left\|O_p-O_{p'}\right\|=\frac{\left(o_{kp}-o_{kp'}\right)\left[o_{kp}\left(1-o_{kp}\right)o_{jp}-o_{kp'}\left(1-o_{kp'}\right)o_{jp'}\right]}{\left\|O_p-O_{p'}\right\|}, \tag{5}$$

$$\frac{\partial}{\partial w_{ji}}\left\|O_p-O_{p'}\right\|=\frac{1}{\left\|O_p-O_{p'}\right\|}\left\{\sum_{k=1}^{2}\left(o_{kp}-o_{kp'}\right)\left[o_{kp}\left(1-o_{kp}\right)w_{kj}o_{jp}\left(1-o_{jp}\right)x_{ip}-o_{kp'}\left(1-o_{kp'}\right)w_{kj}o_{jp'}\left(1-o_{jp'}\right)x_{ip'}\right]\right\}. \tag{6}$$

为保证降维映射中的拓扑信息与原维数空间的一致性，还有非线性方差存储(Non-Linear Variance Conserving, NLVC)映射法[7]和等分正交映射(Equalized Orthogonal Mapping, EOM)[8]来定义 ANN 中不同的目标函数. NLVC 定义了一个目标函数来存储所有维数下数据的总平均方差，该法虽然有较高的计算效率，但也有不足：若由不同的随机种子数给 ANN 设定一个不同的随机产生的初始权值，则对相同的输入，却产生了完全不同的映射结果. EOM 在学习过程中用目标函数来约束输出的协方差矩阵元素值，这保证了降维模型中的变量互相垂直，即不存在线性相关. 与 NLVC 一样，EOM 约束仍然是总体方差存储，所以有相同的不足. 而 CCM 考虑的是个体方差，与 NLVC 和 EOM 相比，CCM 是一种更直观的方法，且对人工神经网络的初始权值的随机性依赖更小. 但与 NLVC 和 EOM 的计算高效性相反，CCM 学习规则(由前述方程给出)的复杂性降低了算法的优化速率，但 CCM 算法和 LVQ(Learning Vector Quantization)组合方法仍可处理大规模问题[9].

2.3 应用聚类分析获得可行域信息

平方和法[4]可将 N 个可行点分割为 G 个类，聚类 g 的矩心为 $Z_{2-D,\,g}$:

$$Z_{2-\mathrm{D},\,gk}=\frac{1}{N_g}\sum_{p=1}^{N_g}o_{gpk}\qquad(k=1,2)\,. \tag{7}$$

类 g 的矩心平方和为

$$S_g=\sum_{p=1}^{N_g}\left\|O_{gp}-Z_{2-\mathrm{D},\,g}\right\|^2\,. \tag{8}$$

平方和法的目标就是要找到一个分割，使

$$S=\sum_{g=1}^{G}S_g\,, \tag{9}$$

最小.

点集 O_g 定义的类 g 满足

$$\sum_{k=1}^{2}(o_k-z_{2-\mathrm{D},\,gk})^2\le r_{2-\mathrm{D},\,g}^2\qquad(k=1,2)\,, \tag{10}$$

其中圆半径

$$r_{2-\mathrm{D},g}=\max\left\{\left\|O_{gp}-Z_{2-\mathrm{D},g}\right\|\right\}\ \ (p=1,2,\cdots,N_g)\,. \tag{11}$$

式(11)确保由式(10)表述的类 g 能够覆盖 O_g 中的所有点. 但是在仅仅为了覆盖一个离中心很远的点而增大类半径的情况下，用式(11)来定义半径并不合适. 因此，本文又提出了一种统计方法，聚类分析的目标就是从概率的角度来确定每一点从属于每一类的可能性. 每一类都符合不同的正态分布，即有不同的平均数(μ)和标准偏差(σ). 如在聚类 g 中，在 O_g 中的点在第 k 维坐标下符合正态分布：

$$f(o_{gpk},\mu_{gk},\sigma_{gk})=\left(\sqrt{2\pi}\sigma_{gk}\right)^{-1}e^{-(o_{gpk}-\mu_{gk})^2/(2\sigma_{gk}^2)}\qquad(k=1,2)\,, \tag{12}$$

这里

$$\mu_{gk}=Z_{gk}=\frac{1}{N_g}\sum_{p=1}^{N_g}o_{gpk}\qquad(k=1,2)\,, \tag{13}$$

$$\sigma_{gk}^2=\frac{1}{N_g}\sum_{p=1}^{N_g}(o_{gpk}-\mu_{gk})^2\qquad(k=1,2)\,. \tag{14}$$

在每一聚类中，根据变量的正态分布可以定义半径为

$$r_{2-\mathrm{D},gk}=\max(\varepsilon_k)\sigma_{gk},\qquad(k=1,2)\,, \tag{15}$$

其中ε 是调节参数，默认值为 1. r 值越大，可排除的可行域越少(即可行域范围越大)，但在接下来的优化过程中所需的计算工作量越多. ε 用于在计算工作量与可行点覆盖之间取得平衡，这个平衡

自动确定很困难，但是由用户通过 2 维映射可视化来参与估计该算法的(中间)结果并得到该平衡值是很容易的.

2.4 将 2 维空间的可行域信息映射回 n 维空间

将在 2 维中确定的聚类信息转换回初始高维 n 维空间：

$$\sum_{i=1}^{n}(x_i - Z_{n-\mathrm{D},gi})^2 \le r_{n-\mathrm{D},gi}^2 \quad (g=1,2,\cdots,G,\ i=1,2,\cdots,n), \tag{16}$$

$$Z_{n-D,gi} = \frac{1}{N_g}\sum_{p=1}^{N_g} x_{gpi} \quad (i=1,2,\cdots,n), \tag{17}$$

$$r_{n-\mathrm{D},gi} = \max(\varepsilon_k)\sigma_{gi} \quad (i=1,2,\cdots,n,\quad k=1,2). \tag{18}$$

3 增强遗传算法

以上描述的聚类分析结果是一类 G 的集. 如式(16)，每一类都定义了一个可行域. GA 的约束处理设计以下适值函数：

$$F(\vec{x}) = \begin{cases} f(\vec{x}) & \text{如果是可行点} \\ \text{很大的值, 如} 10^{20} & \text{如果是非可行点} \end{cases}. \tag{19}$$

为了利用从聚类分析得到的信息，遗传因子按如下方式改进.

3.1 初始化

在初始种群中生成个体来代表如上定义的类. 给定种群大小 N_p，N_p/G(其中 G 由 2.3 节所述确定)点在由每一类定义的区域中随机生成. 在每一类中，每一个体都在其可行域内随机地生成：

$$\left(Z_{n-\mathrm{D},gi} - \max(\varepsilon_k)\sigma_{gi},\ Z_{n-\mathrm{D},gi} + \max(\varepsilon_k)\sigma_{gi}\right) \quad (k=1,2). \tag{20}$$

初始化过程确保了可行域被覆盖和初始种群的多样性. 但可行域由聚类分析确定，可能也覆盖了搜索空间中的不可行区域，即同时生成了不可行点. 不过，与单纯用随机初始化过程而不用可行空间信息的方法相比，该法生成可行点的概率要高.

3.2 变异

变异算子基于参数的变异(Parameter-Based Mutation operator, PBM)算子[10]. 改进变异算子确使变异的结果仍被所获得的类所覆盖：

(1) 找到个体所在的类

(2) 对每一个变量，首先用式(20)来确定大小边界，然后运行 PBM. 因为变异是在由选择出的个体所在的类中进行的，所以其后代也在该类内.

3.3 交叉

应用一种基于实际基因组编码的模拟二进制交叉算子(Simulated Binary Crossover operator, SBC)[11]. 为了利用可行域信息，对该算子添加如下步骤：

(1) 在选出的父代上应用 SBC.

(2) 如果孩子不属于已有聚类，则在该聚类中找到离它最近的点，并将变异算子作用在该点上重新生成一个子体.

3.4 其它 GA 参数

应用锦标赛选择算法，种群更新策略是先添加新生成的子代到种群中，再在生成下一代之前将种群中最差的个体删去.

当最优解不发生改变，N_{conv}=10 时认为达到收敛. 交叉率 r_c=0.9，变异率 r_m=0.1.

4 简 例

为了研究提出的 IGA 的优化效率，选择 2 维的约束最小化问题.

$$\begin{aligned}&\text{Minimize} \quad \left(x_0^2+x_1-11\right)^2+\left(x_0+x_1^2-7\right)^2\\&\text{Subject to} \quad g_1: 4.84-\left(x_0-0.05\right)^2-\left(x_1-2.5\right)^2\le 0 \quad (0\le x_i\le 6,\ \ i=0,1)\,.\\&\qquad\qquad g_2: x_0^2+\left(x_1-2.5\right)^2-4.84\ge 0\end{aligned} \tag{21}$$

无约束目标函数的最优解在(3,2)处，此时函数值有最小值 0；约束的最优解为(2.247, 2.382)，此时函数值为 13.591，可行域是狭窄新月形.

对种群进行 10 次不同的初始化后，用改进的遗传算法进行优化. 将最优、平均和最差优化结果连同收敛代数在表 1 列出. 表中有 5 个不同的例子：在(A), (C)和(D)中，分别用有 3, 4, 5 个类的可行域信息的 IGA 算法，且得到的信息都有用户参与；而在(B)中，可行域信息有 4 个类，但没有经过用户的可视化调整. 在(E)中，应用了约束处理方法[12].

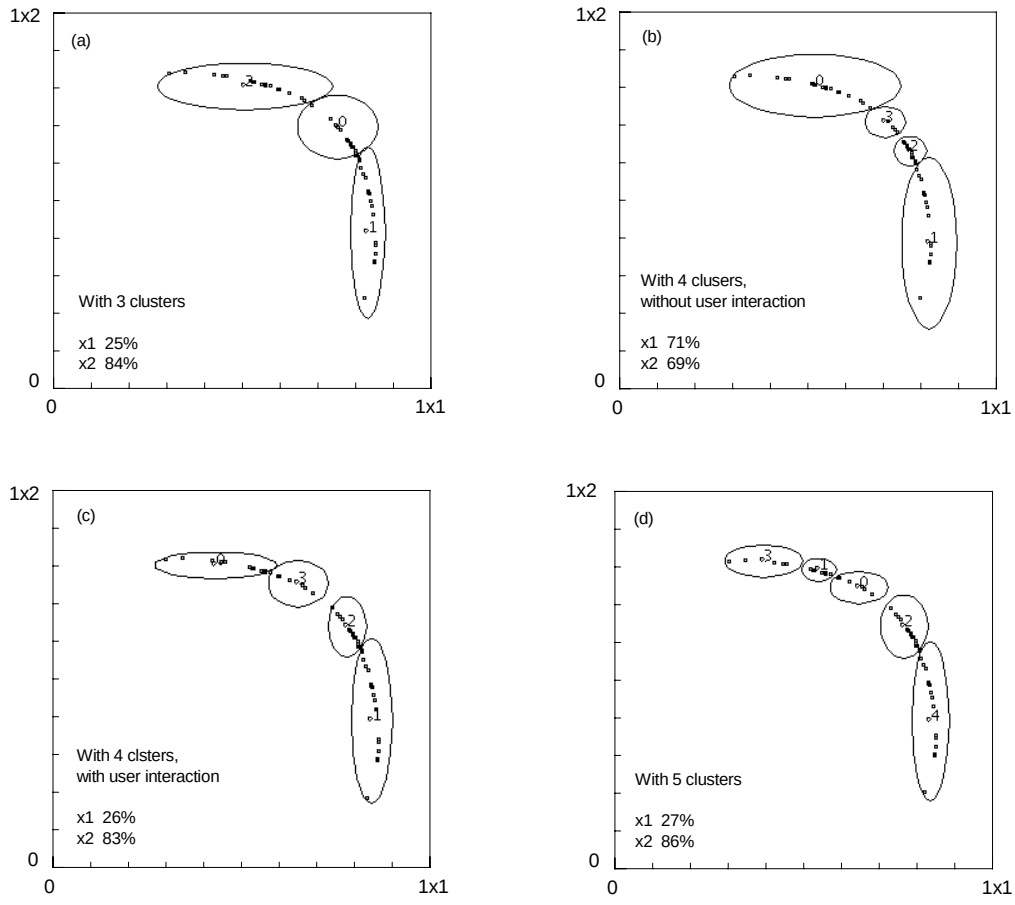

图 2 CCM 映射聚类分析

Fig.2 The cluster analysis based on CCM

表 1 GA 结果比较

Table 1　Comparison of GA results

Different cases	Optimal value			Terminate generation
	Best	Medium	Worst	
(A) IGA+3	13.6110	40.0910	250.1120	30
(B) IGA + 4 without user interaction	13.5913	24.0743	172.8100	33
(C) IGA+4	13.5908	16.4210	172.8100	33
(D) IGA+5	13.5908	15.4530	117.0200	35
(E) DebGA	13.5908	13.61673	114.6900	40

通过比较例(A), (C)和(D)中 GA 的性能，可以说明类数目越多，就可利用 GA 得到越好的最优解. 这是因为类数目越多的信息覆盖较少的非可行域，这使 GA 在可行域内高效进化. 但也可看出，聚类数目越多，所需计算量越大. 所以需要借助用户的可视交互作用来确定合适的类数目和对可行域信息进行调整. 本例如图 2 所示，由用户可视化确定较优的聚类数为 4. 比较例(B)和(C)的 GA 性能，可以看出在用户参与下，聚类的中心和半径发生了变化，IGA 的性能也随着聚类信息的改进得到了提高. 比较例(C)和(D)的结果，两者都找到了最优解. 但在例(D)中，需要更大的计算工作量来计算约束偏离值，例(C)中则不用. 且对于工业实际应用问题，约束偏离值有时很难得到.

5　结 语

本文首先提出了一种基于前向人工神经网络(ANN)的聚类约束映射(CCM)进行从初始 n 维空间到 2 维的维数缩减映射，同时在缩减的维空间中保存聚类信息. 然后用作用于 2 维空间的凝聚算法[6]进行聚类分析. 最后，将从 2 维空间得到的聚类信息映射回 n 维空间，从而得到在初始维数下的可行域信息，该信息最后用于改进的遗传算法 IGA. 该法保证种群在可行域中进化，且不需要任何惩罚参数.

通常情况下，交叉率和变异率对不同的世代数要进行适当调节[1]：在开始进化时，由于种群有较高的多样性，采用较高的交叉率可使算法用较少的计算量找到局部最优解；经过几代进化后，种群多样性随着个体陷入局部最优而明显下降，此时应降低交叉率，提高变异率来保持其多样性. 如何在进化过程中适当地调节交叉率和变异率，如何使用户通过模拟程序来直接监视种群的多样性，是今后的工作方向.

符号表：

E	距离比率值方差	k	原始数据缩减到的维数下标 (k=1,2)
K	原始数据缩减到的维数	N	可行点数目
N_{conv}	最优值未发生改变时允许的最大世代数	N_g	聚类 g 的可行点总数目
N_{p}	种群大小	O	输出 2 维矢量(或点)
O_g	聚类 g 的点集	O_{gp}	聚类 g 的第 p 个点
o_{gpk}	聚类 g 中第 p 个点的第 k 维坐标点 (g=1,⋯,G; p=1,⋯N_g; k=1,2)	o_k	输出矢量 O 的第 k 维坐标
		p, p’	节点下标
r_{c}	交叉率	r_{m}	变异率
w_{kj}	前向 ANN 输出层 k 和隐藏层 j 之间的权值	w_{ji}	前向 ANN 隐藏层 j 和输入层 i 之间的权值
X	给定的 n 维矢量(或点)	X_g	被映射到 2 维聚类 g 的初始 n 维空间点
x_i	给定矢量 X 的第 i 维坐标	(x_i^l, x_i^u)	x_i 的变化范围
$Z_{2\text{-D},\,g}$	输出 2 维空间下聚类 g 的矩心	$Z_{2\text{-D},\,gk}$	输出 2 维空间下聚类 g 各维的矩心
$Z_{n\text{-D},\,g}$	初始 n 维空间下聚类 g 的矩心	$Z_{n\text{-D},\,gi}$	初始 n 维空间下聚类 g 各维的矩心
$\|\cdot\|$	两模式点间的欧几里得距离	μ	正态分布平均数
σ	正态分布标准偏差	ε	计算工作量与可行点覆盖间取得平衡的调节参数，默认值 1

参考文献：

[1] Goldberg D E. Genetic Algorithms in Search Optimization and Machine Learning [M]. Reading: Addison Wesley, 1989.

[2] Michalewicz Z. Genetic Algorithms, Numerical Optimization, and Constraints [A]. Eshelman L. Proceedings of the Sixth International Conference on Genetic Algorithms [C]. San Mateo: Morgan Kauffman, 1995. 151−158.

[3] Michalewicz Z, Schoenauer M. Evolutionary Algorithms for Constrained Parameter Optimization Problems [J]. Evol. Comput., 1996, 4(1): 1−32.

[4] Gordon A D. Classification [M]. New York: Chapman and Hall, 1981.

[5] Kohonen T. Self-organizing Maps [M]. Berlin, Heidelberg: Springer, 1995.

[6] Ward J H. Hierarchical Grouping to Optimize a Objective Function. [J]. J. Am. Statist. Assoc., 1963, 58: 236−244.

[7] Pao Y H. Dimension Reduction, Feature Extraction and Interpretation of Data with Network Computing [J]. Int. J. Pattern Recogn. Artif. Intell., 1996, 1(5): 521−535.

[8] Zhou Meng. Visualization and Self-organization of Multidimensional Data as an Aid to Understanding [M]. Electrical Engineering and Applied Physics, 1999.

[9] Wang K, Fraga E, Salhi A. A Cluster Identification Using Parallel Coordinate System and Genetic Algorithm for Knowledge Discovery and Nonlinear Optimization [A]. European Symposium on Computer Aided Process Engineering-12 (ESCAPE 12) [C]. Hague, 2002. 1003−1008.

[10] Deb K, Goyal M. A Combined Genetic Adaptive Search (GeneAS) for Engineering Design [J]. Comput. Sci. Informa., 1996, 26(4): 30−45.

[11] Deb K, Agrawal R B. Simulated Binary Crossover for Continuous Search Space [J]. Compl. Sys., 1995, 9: 115−148.

[12] Deb K. An Efficient Constraint Handing Method for Genetic Algorithms [J]. Comput. Method Appl. Mech. Eng., 2000, 186: 311−338.

Cluster Analysis and Visualization Enhanced Genetic Algorithm —I. Eduction, Principle and Analysis

WANG Ke-feng, SUN Xiao-jing, YAO Ping-jing

(*1. Dept. Chem. Process System Eng., Dalian University of Technology, Dalian, Liaoning 116012, China*)

Abstract: Genetic Algorithms (GA) based on penalty function methods have been the most popular approach to constrained optimization problems because of their simplicity and ease of implementation. But how to find appropriate penalty parameters needed to guide the search towards the constrained optimum in the penalty function approaches is very difficult. A new cluster analysis based on visualization is proposed to address the constrained optimization problems. First, a Cluster Constrained Mapping (CCM) method based on feed-forward Artificial Neural Network (ANN) is proposed for dimension-reduction mapping from the original *n*-D space to 2-D, conserving the cluster information in the reduced dimensional space. Then the agglomerative algorithm that works in 2-D space is called upon for cluster analysis. Its parameters are provided through visualization and subsequent interaction with the user. Finally, the cluster information is derived from 2-D back into *n*-D to obtain the feasible region knowledge in the original dimensions, which is used in the IGA. The enhanced GA, incorporating a new cluster analysis method through data visualization (CCM) and user interaction guarantees the process of evolution in feasible regions without requiring any penalty parameters.

Key words: visualization; cluster analysis; constrained optimization; genetic algorithm