基于小生境技术和聚类分析的人工免疫算法*)

郝晓丽　谢克明
(太原理工大学计算机学院　太原 030024)

摘　要　针对标准人工免疫算法存在的早熟收敛和后期收敛速度慢的问题，本文提出了一种基于小生境技术和聚类分析的改进的人工免疫算法。首先运用嵌入进化标记的小生境技术对初始种群进化，“排挤机制”有效地保持种群的多样性，防止了早熟，而标记种群的进化方向则加快了算法的收敛速度。其次聚类方法的应用使得在各极值点附近形成了聚类区域，在不同的聚类区域运用人工免疫的趋同算子和异化算子分别进行粗搜索和细搜索，以保证全局寻优的速度和精度。仿真结果表明，该改进算法较之标准免疫算法，有更快的收敛速度、更强的全局搜索能力和更好的寻优精度。

关键词　人工免疫算法，小生境，聚类，算子

Artificial Immune Algorithm Based on Niche Technology and Cluster Analysis

HAO Xiao-Li　XIE Ke-Ming
(College of Computer and Software Engineering, Taiyuan Technology University, Taiyuan 030024)

Abstract　Due to premature convergence and low speed of latter convergence in conventional artificial immune algorithm, the new method is introduced in the paper which is improved artificial immune algorithm based on niche technique and cluster analysis. Firstly niche technique with revolutionary recording is taken to initial population. “exclusion mechanism” can maintain population diversity to avoid premature, and labeling evolution direction dynamically can improve convergence speed. Then cluster analysis is applied to obtain cluster areas nearby extremums. Different operators are taken in different areas respectively. Similar-taxis operator is employed to realize optimization within cluster areas, while dissimilation operator between them. Parallel searching in coarser and finer layer can ensure the speed and precision of global optimization. Simulation shows that the improved algorithm has higher convergence speed, better capability of global searching and better optimization precision.

Keywords　Artificial immune algorithm, Niche, Cluster, Operator

1　引言

人工免疫算法是模拟自然免疫系统功能的一种智能方法，它通过学习外界物质的自然防御机理，提供了噪声忍耐、无师学习、自组织、记忆等进化学习机理[1,2]，是自然免疫系统在进化计算中的一个应用。自产生以来，研究工作者提出了许多改进的人工免疫算法，如免疫规划算法，通过引入“免疫算子”-接种疫苗和免疫选择，模拟人体免疫系统所特有的自适应特性，加快了算法收敛的速度，但在用其求解多峰函数的最优解时，无法同时找到所有的全局最优解，甚至容易陷入局部最优。再如基于信息熵的免疫算法，通过抗体基因位的信息熵求出抗体的浓度并基于浓度进行调节操作，使抗体不断优化，最终找到最佳抗体，即最优解。但这种方法的不足之处在于因计算复杂和含有冗余的计算信息从而导致算法收敛速度慢。由此可以看出，一些改进的免疫算法虽然理论上可以搜索到全局最优，但依然存在三个严重的缺陷，(1)容易陷入局部最优的平衡态而导致早熟收敛，(2)由于进化后期搜索停滞不前而导致后期收敛速度慢，(3)由于初始种群和后期进化存在一定的随机性，使得搜索的结果精度不高。

因此，提高免疫算法的优化效果需要考虑的因素应包括：增强算法的全局搜索能力；增加种群的多样性，避免早熟；避免局部极值区域存在过多的个体，影响收敛速度等。鉴于小生境是一种防止早熟，提高搜索速度和精度的有效方法[3]，本文提出了一种基于小生境技术和聚类分析的改进免疫算法。首先将小生境技术应用于初始种群的进化，即基于“排挤机制”[4]，通过不断度量个体之间的相似性用以限制相似个体的数量，随着排挤过程的进行，群体中的个体逐步被分类，从而形成各个小的生存环境，有效地维持了群体的多样性。当进化到一定程度之后，再进行聚类分析[5]，由此可以获得分布在各个极值点附近的聚类区域，在各个聚类中心处，利用人工免疫算子进行搜索从而获得极值点，其余个体按照小生境技术在聚类区域外进一步进化。在进化的过程中，该算法还为每个个体标记了进化方向，即以每个个体为初始点，按标记的进化方向局部寻优，提高了搜索的速度和精度。仿真结果表明，这种算法能够有效地防止早熟收敛，还可以明显增强算法的全局搜索能力，极大提高了免疫算法的收敛速度，有利于并行实现。

2　标记进化方向的小生境技术

在免疫算法中引入小生境思想，让种群中的个体不是聚集在一种环境中，而在不同特定的生存环境中进化。这样可以使算法在整个解空间中搜索，以找到更多的最优个体，避免在进化后期适应度高的个体大量繁殖，充斥整个解空间，导致算法停止在局部最优解上。

本文建立的小生境模型可标记个体的进化方向，其特点：(1)“基于排挤机制”，在进化过程中，能不断排挤掉相似的个体，有效地保持种群的多样性，从而搜索到更多的解。(2)利用进化过程中的有用信息，为每个个体标记进化方向，即以每个个体为初始点，按标记的进化方向继续局部寻优，以提高解的精度。

2.1　小生境实现原理

将小生境的“排挤机制”引入免疫算法，是在进化过程中，

*)基金项目：国家自然科学基金(60374029)；山西省回国留学人员基金(2004-18)。**郝晓丽**　博士生，主要研究方向为人工智能，进化计算，故障检测等；**谢克明**　教授，博士生导师，主要研究方向为人工智能等。

通过比较个体间的相在似性，抑制相似个体的生长，避免大量个体趋向于同一，从而保持了种群多样性。其实现方法描述如下：记第 i 个个体为 P_i，第 j 个个体为 P_j，$\| P_i - P_j \|$ 为 P_i，P_j 间的欧氏距离，L 为一较小的正数，$f(\cdot)$为适应度函数。比较种群中任意两个个体的相似性，若 $\| P_i - P_j \| \leqslant L$，表明 P_i 与 P_j 相似程度较大，则对适应度较小者施加一个较强的惩罚函数，即 $f(P_j) = penalty * f(p_j)$，使其适应值变得极小，在以后的进化中 P_j 会以极大的概率被淘汰掉。根据排挤原理，这样可以实现小生境的进化环境。

2.2 进化方向的标记

个体进化方向的标记可在上面的小生境实现中同步完成。其描述如下：

设单变量函数 $y=g(x)$，$x_1, x_2 \in A$ 且 $|x_1 - x_2| < \rho$，ρ 为一较小的正数。设目标函数 $J = \max\limits_{x \in A} g(x)$，由图 1 可知，$x_1$ 比 x_2 更优。

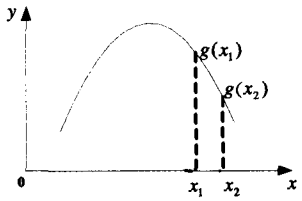

图 1 进化方向原理说明图

可以根据 x_1 与 x_2，$g(x_1)$与 $g(x_2)$的大小关系，定出 x_1（更优者）的进化方向：由 $x_1 < x_2$，$g(x_1) > g(x_2)$及 $|x_1 - x_2| < \rho$，则在 x_1 的一个邻域内 $g(x)$是下降的，可以推断出，存在一个很小的正数 ε，使得 $g(x_1 - \varepsilon) > g(x_1)$，即 $(x_1 - \varepsilon)$是比 x_1 更优的点，因此 x_1 的进化方向为 -1。

由此，可设 x_{ik} 为个体 P_i 的第 k 个决策变量，x_{jk} 为个体 P_j 的第 k 个决策变量。当 $\| P_i - P_j \| < L$ 时，可以通过比较 $f(P_i)$与 $f(P_j)$，x_{ik} 与 x_{jk} 的大小关系确定出 P_i 与 P_j 中更优者的进化方向分量 $d_k (k=1,2,\cdots,n)$。

当 $f(P_i) > f(P_j)$时（若 P_i 为更优者），P_i 的进化方向为：

$$d_k = \begin{cases} -1 & x_{ik} < x_{jk} \\ 0 & x_{ik} = x_{jk} \\ 1 & x_{ik} > x_{jk} \end{cases}$$

当 $f(P_i) < f(P_j)$时（若 P_j 为更优者），P_j 的进化方向为：

$$d_k = \begin{cases} -1 & x_{ik} > x_{jk} \\ 0 & x_{ik} = x_{jk} \\ 1 & x_{ik} < x_{jk} \end{cases}$$

将进化方向的标记方法融入小生境思想中，设计出能标记进化方向的小生境免疫算法，加快算法的收敛速度。

3 基于小生境技术和聚类分析的人工免疫算法

算法的基本思想：群体首先按照嵌入进化标记的小生境算法进化若干代，当形成聚集趋势后，对群体进行聚类分析。然后运用人工免疫算子，即选择、趋同和异化算子对类内、类外抗体进行操作。在每个类内，通过趋同算子使类内的抗体相互竞争，从而快速收敛到相应的局部极优点。在不同类之间，通过异化算子，交换两个类之间的最优抗体所携带的最优信息，从而打破类内部的平衡状态，从而能快速收敛到全局最优点。同时，由于聚类操作使得类外个体数量减少，为了在类外进行有效搜索，因此及时补充类外个体数量，使其保持固定规模，然后对各个类外抗体分别按照小生境算法进化，再进行免疫算子操作。如此循环直至找到全局最优抗体。算法描述如下：

```
设种群规模为 M，类的个数为 C
随机产生 M 个抗体构成初始种群，计算每个抗体的适应值
While (不满足停止条件)
{
  If(第一次聚类前)执行若干代嵌入进化标记的小生境算法
     If(满足聚类条件)
     {
       聚类；
       If(某类、外抗体数偏少)；
         {补充类内个体；
          补充类外个体；}
       Else
     {For (1 到 C 类)
     对类内、类外抗体进行人工免疫算子操作
      }
     }
}
```

对上述算法，说明如下：

(1)当运用小生境技术对种群进化，且进化到一定程度时，再进行聚类分析。在各个聚类中心处，利用人工免疫算子进行局部搜索从而获得极值点。聚类方法采用最邻近规则的试控法并动态调整聚类中心，算法描述如下：

```
For (i=1 到 C) 聚类中心 Ci=NULL;
取个体 X1 作为聚类中心 C1
For (i=2 到 n)
{
  for (j=1 到 C)
      if (Cj 存在且 Xi 与 Cj 距离小于阈值)
      {Xi,Cj 同属一类，调整聚类中心 }
}
```

根据上述算法，可以获得一个聚类结果。其中，聚类前对抗体按适应度大小降序排序，聚类时按此顺序分别取个体作为聚类中心，这将确保在当前最优抗体附近有一个聚类区域，聚类后若类中个体小于规定数量，则对类内最优抗体进行复制，通过这种方法来补充类内抗体，可以保证每个小生境有足够的抗体进行搜索。

(2)人工免疫算法中的算子

当完成聚类，每个类都可设为一个初始种群 $A(k)$。

选择算子：根据适应度函数，计算 $A(k)$中每个抗体的适应度，将其中一些适应度最高的抗体作为免疫记忆抗体保留，得到群体 $B(k)$。

趋同算子：群体 $B(k)$中的抗体为成为胜者而竞争的过程叫趋同。在趋同半径内进行趋同操作，形成群体 $C(k)$。

异化算子：类内的种群经过一定时间的进化容易陷入平衡状态，在该算法中，为打破这种状态，每进化一代，都需要选取每个类内种群中通过趋同算法得到的最优抗体，与另一类内种群的最优抗体进行竞争。通过异化算子，将交换两个子种群的最优抗体所携带的最优信息，以打破类内部的平衡态，由此产生种群 $D(k)$。

4 仿真实验分析

为了验证该算法的寻优能力和收敛速度，本文用离线性能指标来体现算法运行过程中各进化代的最佳性能值的累积平均，由此可以反映算法的收敛能力，并且在进化过程中每进化一代就统计目前各代中最佳适应度值，计算进化代数的平均值。为了便于比较，本文对两个测试函数分别采用标准的人工免疫算法（AIA—Artificial Immune Algorithm）和基于小生境和聚类分析的人工免疫算法（NCAIA—Artificial Immune Algorithm Based on Niche and Cluster Analysis）进行寻优，两个算法都具有相同的初始种群，终止代数为 100。

(1)Branin rcos 函数是一个全局极小值测实函数，它有 3

个全局极小值。该函数的定义是：

$$f(x_1,x_2)=a\cdot(x_2-b\cdot x_1^2+c\cdot x_1-d)^2+e\cdot(1-g)$$

其中，$a=1,b=5.1/(4.\pi^2),c=5/\pi,d=6,e=10,g=1/(8\cdot\pi),x_1\in[-5,10],x_2\in[0,15]$。

函数的全局极小小值 $f_{\min}(x_1,x_2)=0.39788$。相应的 $(x_1,x_2)=(-\pi,12.275),(\pi,2.275),(9.42478,2.475)$。

该函数的三维图像如图 2 所示。

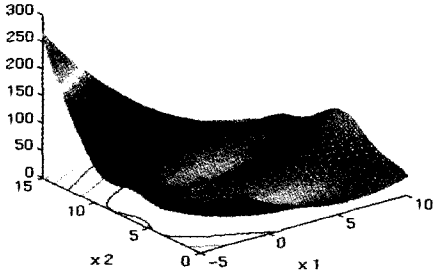

图 2　Branin rcos 函数图像

具体优化数值结果如表 1 所示。

表 1　Branin rcos 函数优化结果

算法	最优解	自变量 x_1	自变量 x_2
NCAIA	0.39788742933166	3.14147174853691	2.27505660893729
	0.39788760279687	9.42496806187726	2.47489284357722
	0.39789017358823	−3.14092929544622	12.27424440798476
AIA	0.39788738972704	3.14151331397734	2.27510401404908
	0.39788804981837	9.42443595516518	2.47435025132843
	0.39788914995162	−3.14218496260581	12.27675193768933

优化比较结果如图 3，图 4 所示。

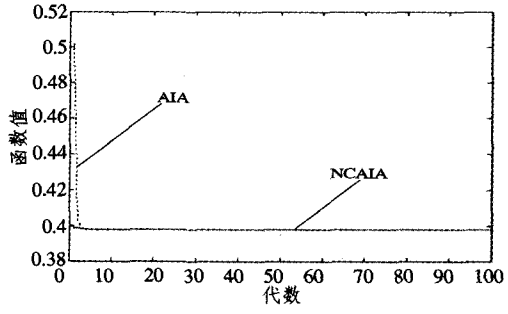

图 3　Branin rcos 函数的寻优比较

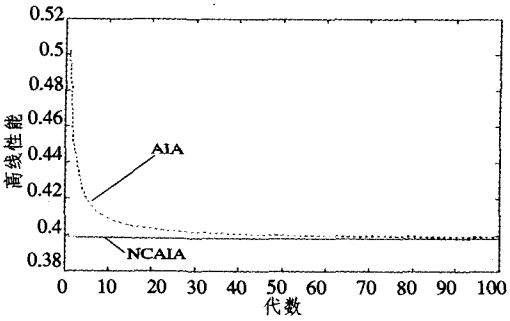

图 4　Branin rcos 函数优化中算法的离线性能比较

从图 3 的比较中可以看到，在此类函数的优化中，两个算法都是有效的，即能够找到最优解，而从收敛速度来看，NCAIA 明显较 AIA 快。从图 4 可以看出 NCAIA 的离线性能比 AIA 好，虽然函数有许多局部极小值，但是 AIMEA 仍能较快地找出全局最优解，可见其有效地避免了“早熟”现象，不容易陷入局部极值，其收敛速度也要优于 AIA。

(2) Schaffer 函数是一个全局极小值测试函数，它有 1 个全局极小值。但是有多个局部最小值。它的强烈震荡性质及它的全局最优点被次最优点所包围的特性，使得很难找到最优解。该函数的定义是：

$$f(x_1,x_2)=\frac{\sin^2(\sqrt{x_1^2+x_2^2})-0.5}{[1.0+0.001\cdot(x_1^2+x_2^2)]^2}+0.5$$

其中，$x_{1,2}\in[-10,10]$。该函数的全局极小值 $f_{\min}(x_1,x_2)=0$，相应的 $x_i=0,i=1,2$。函数图像如图 5 所示。

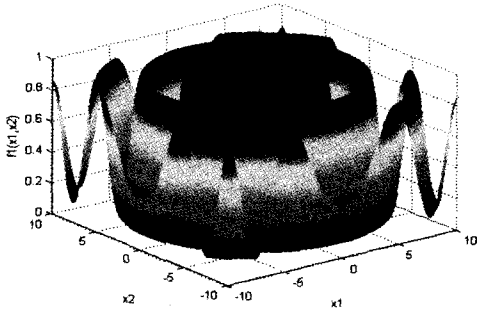

图 5　Schaffer 函数图像

具体优化数值结果如表 2 所示。

表 2　Schaffer 函数优化结果

	NCAIA	AIA
最优解	$1.19348975147204\times10^{-14}$	0.00971591003241
自变量 x_1	$0.10864090414429\times10^{-6}$	1.82648966766530
自变量 x_2	$0.01207741941766\times10^{-6}$	2.55224286010890

优化比较结果如图 6，图 7 所示。

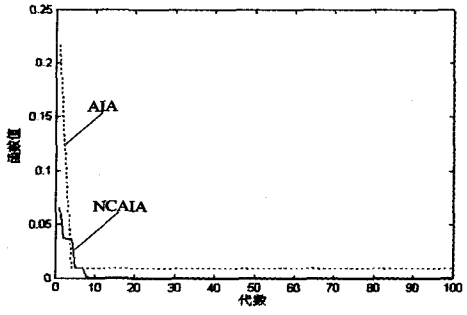

图 6　Schaffer 函数的寻优比较

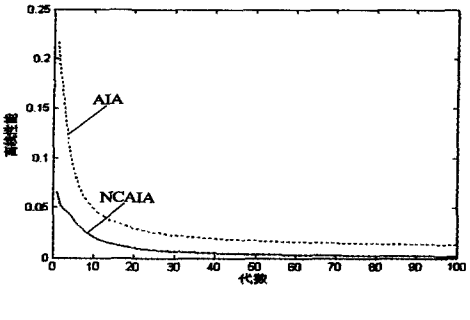

图 7　Schaffer 函数优化中算法的离线性能比较

从图 6 和图 7 的比较中可以看到，在此类函数的优化中，NCAIA 算法能够很容易地从局部最优解中跳出，更快更精确地找到最优解，而 AIA 算法的收敛速度和寻找最优解的精度都不是十分理想。

结论 针对标准人工免疫算法中存在的早期收敛和后期收敛速度慢的问题，提出了一种基于小生境技术和聚类分析的人工免疫算法。利用嵌入进化标记的小生境技术，不仅加快了算法的收敛速度，而且保持了种群的多样性，有效地防止早熟。当种群进化到一定程度后，进行了聚类分析，使得人工免疫算法的异化算子在类与类之间可以进行全局搜索，即运用粗搜索找到适应度较高的区域，再用趋同算子在类内抗体中进行局部搜索以寻求高精度的解。算法通过粗和细的两层搜索，保证了全局寻优的速度和精度。实验表明，该算法是一种寻优能力强、效率和可靠性更高的优化算法，其综合性能比标准的人工免疫算法有了一定的提高。

参 考 文 献

1 王磊，潘进，焦李成. 免疫算法[J]. 电子学报，2000，28(7)：74～78
2 Vargas P A, Castro L N, Von Zuben F J. Artificial immune systems as complex adaptive systems. In: Proceedings of International Conference on Artificial immune systems, 2002, 1: 115～123
3 Miller B L, Shaw M J. Genetic Algorithms with Dynamic Niche Sharing for multimodal Function Optimization. In: IEEE International Conference on Evolutionary Computation, Piscataway, NJ: IEEE Press, 1996, 786～791
4 Goldbergde, Richardson J. Genetic Algorithm with Sharing for Multimodal Function Optimization[A]. In: Proceedings of 2nd International Conference on Genetic Algorithm, Lawrence Erlbaum Associates[C], 1987. 41～49
5 Jain A K, Dubes R C. Algorithms for clustering[M]. Englewood Cliffs: N J Prentice Hall, 1988

(上接第 134 页)

们对待离群点的正常思维方式。对于线性不可分的数据样本集，本文提出的离群聚类算法是在引入非线性函数，映射至特征空间进行聚类以及根据等权值线标识离群点的结果，所以能够取得较好的聚类效果。另外，由于本文提出的算法加入了权值的概念，因此能够更容易地发现样本数据集中的离群点。

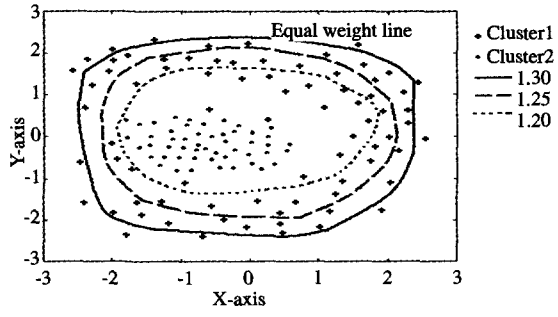

图 6 本文提出的算法对线性不可分样本数据集聚类的结果

对实验结果分析发现：

①实验结果与样本数据点的个数和权值有很大关系，选取不同数目的样本数据点和分配所得的不同权值对边缘点的影响都会比较大；由于离群点通常远离任何一个类中心，从而分配一个大权值 w_k($1/w_k^q$ 小)。权值指数 q 等于 1.5 时，本算法就是基于重心的算法，而权值指数 q 趋近 0 时，样本数据点取所有点时，本算法就是基于所有样本数据点的算法。经过测试权值指数 q 选取 1 时效果较好。②因为测试数据选取有限，所以没能进行更广泛的测试，通过观察分类结果，发现其中有很多是分散样本数据点归为一类，而一类最多包含的样本数据又过多，可见每一类所包含文本数不太平衡，说明在许多方面还有待改进。

结束语 本文在基于核的 PP 主成分的基础上提出一种离群聚类算法。该算法首先用数据变换对输入数据进行预处理，数据预处理的一个重要结果就是得到降维的投影向量。通过引入非线性变换函数，把预处理所得数据映射到特征空间，通过为特征空间中的向量分配一个动态的权值，最终在优化经典的 FCM 目标优化迭代函数基础上，借助权值发现样本集中的离群点，特别是对于一些线性不可分的数据集，在运用传统算法失败的情况下，本文提出的离群聚类算法仍然能在取得良好的聚类效果的同时发现其离群点。

该算法优点包括：

a 能够有效地解决高维空间中数据的稀疏问题；

b 能够找到合适的衡量办法给出高维子空间中离群点的物理意义；

c 适用于线性不可分的数据集；

d 对高维空间中的数据仍然是计算高效的。

最后，我们也证明了离群聚类算法的收敛性。仿真实验表明了该算法的有效性。

参 考 文 献

1 夏火松主编．数据仓库与数据挖掘技术．北京：科学出版社，2004
2 Barnett V, Lewis T. Outliers in Statistical Data[M]. New York: John Wiley and Sons, Inc., 1994
3 Franco P, Ian S M. Computational Geometry: an Introduction [M]. New York: Springer-Verlag, 1988
4 Knorr Edwin M, Ng R T. Algorithms for Mining Distance-Based Outliers in Large Datasets[R]. In: Proceedings of the 24th International Conference on Very Large Data Bases. New York: Morgan Kaufmann, 1998. 392～403
5 Breunig M M, Kriegel H P, Ng R T, et al. LOF: Identifying Density-based Local Outliers[R]. In: Chen W, Naughton J F, Bernstein P A, eds. Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, Texas: ACM Press, 2000. 93～104
6 Spiros P, Hiroyuki K, Gibbons P B. LOCI: Fast Outlier Detection Using the Local Correlation Integral[R]. In: Proceedings of the 19th International Conference on Data Engineering, 2003. 315～326
7 Beyer K, Goldstein J, Ramakri S R, et al. When is nearest neighber meaningful? In: Been, C, Buneman P, eds. Proceedings of the 7th Intemational Conference 0n Data Theory Lecture Notes In Computer Science 1 540. Jernsalem: Spnnger, 1999. 217～235
8 Li Y. Reforming the theory of invariant moments for pattern recogoition[J]. Pattern Recognition, 1992, 25(7): 723～730
9 魏藜，钱卫宁，周傲英. SLOT：基于估计的高效子空间局部离群点发现[J]. 计算机科学，2002，29(8)：122～125
10 Hinneburg A, Aggarwal C C, Keim D A. What is the nearest neighbors in high dimensional space? [n]. In: Abbadl A E, Brodie M L, Chakravarthy S, et al, eds. Proceedings of the 26th International Conference on Very Large Data Bases Cairo: Morgan Kaufmann, 2000. 506～515
11 项静恬，史久恩. 非线性系统中数据处理的统计方法．北京：科学出版社，1997. 171～194
12 Friedman J H, Tukey J W. A projection pursuit algorithm for exploratory data analysis[J]. IEEE Tram Computer, 1974, 23(9): 881～889
13 Suykens J A K, Gestel T V, Vandewalle J, et al. A Support Vector Machine formulation to PCA Analysis and its Kernel version[R]. [ESAT-SCD-SISTA Technical Report 2002-68]. Belgium: Katholieke Universiteit Leuven, 2002
14 Suykens J A K, Gestel T V, Vandewalle J, et al. A Support Vector Machine formulation to PCA Analysis and its Kernel version[R]. [ESAT-SCD-SISTA Technical Report 2002-68]. Belgium: Katholieke Universiteity Leuven, 2002
15 Pal N R, Bezdek J C. On cluster validity for the fuzzy c-mesns model[J]. IEEE Trans Fuzzy System, 1995, 3(3): 370～379