基于小生境遗传算法的分类优化方法*

李隽颖，楼晓俊
（中国科学院上海微系统与信息技术研究所 无线传感器网络与通信重点实验室，上海 200050）

摘　要：对于多分类问题，大多是经二分类器组合进行训练的，在分类类别多、特征维数高时，存在识别准确率不高和训练速度较慢的问题。将超球支持向量机应用到多类问题，为每个类建立一个超球体模型，通过多个超球体划分样本空间。采用改进的基于排挤的小生境遗传算法（improved crowding niche genetic algorithm，ICNGA）进行特征选择，为不同的目标类别寻找最优的特征子集，优化超球支持向量机的输入。利用 UCI 标准数据集的数值实验表明，在分类数据类别较多、特征维数较高时，经过 ICNGA 特征选择之后的多超球支持向量机的识别准确度更好，非常适合解决类别数多、特征维数高的分类问题。

关键词：遗传算法；排挤小生境技术；超球支持向量机；特征选择
中图分类号：TP183　　**文献标志码**：A　　**文章编号**：1001-3695(2012)05-1787-04
doi:10.3969/j.issn.1001-3695.2012.05.049

Classifier optimization method using niche genetic algorithm

LI Jun-ying, LOU Xiao-jun
(*Key Lab of Wireless Sensor Network & Communication, Shanghai Institute of Micro-system & Information Technology, Chinese Academy of Sciences, Shanghai* 200050, *China*)

Abstract: According to multi-classify problem, the multi-classes classifier constructed by binary classes classifier are usually very slow to be trained. When a large number of categories of data are to be classified, the training work could be very difficult. Hyper-sphere support vector machine (HSSVM) can be extended to solve this multi-classification problem. Each category data trains only one HSSVM, the sample space is divided by multiple optimal hyper-spheres. In order to improve the performance of classifier, this paper used improved crowding niche genetic algorithm (ICNGA) to select features, chose optimal feature subset for different target classes. Using UCI data set of numerical experiment shows that the classifiers have a higher accuracy if ICNGA has been used for feature selection, especially the sample data has a large number of categories or feature vectors.

Key words: genetic algorithm(GA); crowding niche technology; hyper-sphere support vector machine(HSSVM); feature selection

0　引言

在模式识别的应用中，标准支持向量机（support vector machine, SVM）是二分类器[1]，由于实际问题大部分是多分类问题，将 SVM 推广到多分类情况，成为了机器学习领域的研究热点。对于多类问题，有一对多（one-against-rest）[2]、一对一（one-against-one）[3]、层树分类[4,5]、多类 SVM[6] 四种常用方法。

用一个半径尽可能小的超球体将训练样本包含其中的思想最早见于文献[7]，Scholkopf 等人将其用于 VC 维的估计。1999 年，Tax 等人[8]进一步提出用超球面作为划分面的模型，正常数据在球面以内，异常数据在球面以外。在求解这个最优化问题之后，对每一点如果判定函数为正就是正常点，反之则为异常点。超球支持向量机（HSSVM）的思想是寻找一个超球体，在包含的一类训练样本尽可能多的同时，使其半径尽可能小。其目标函数为

$$\min_R R^2 + C\sum_{i=1}^{l}\zeta_i \tag{1}$$

$$\text{s.t.}\quad \| \Phi(x_i) - a \|^2 \leqslant R^2 + \zeta_i \tag{2}$$

$$\zeta_i \geqslant 0 \tag{3}$$

其中：R 为球体半径；a 为球心；ζ 为松弛变量；l 为训练样本的个数；C 为正则化参数，控制对错分样本的惩罚程度，实现对球的大小和所包含的样本数之间的折中。与 SVM 类似，当样本点非线性可分时，使用非线性映射 $\Phi(x_i)$，将样本点映射到高维空间，在那里样本是线性可分的，引入核函数 $K(x,y)$ 以及惩罚因子，决策面在特征空间为超球面，通过选择合适的核函数求出的目标样本区域非常紧凑，具有更好的目标识别能力。惩罚参数 C 还可以剔除学习样本中孤立的噪声干扰，这种支持向量数据描述的方法不仅具有支持向量机的优点，而且还能提供待测样本接近目标程度的一个度量。式(1)～(3)的对偶问题为

$$W(a) = \min \sum_i \sum_j \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{l} \alpha_j K(x_i, x_j) \tag{4}$$

其中：

$$\sum_{i=1}^{l} \alpha_i = 1 \tag{5}$$

式(4)为二次规划问题，求解上述问题，即可确定球心和

收稿日期：2011-09-23；**修回日期**：2011-11-02　　**基金项目**：国家重点基础研究发展计划资助项目(2011CB302906)；国家重大科技专项基金资助项目(2010ZX03006-004)

作者简介：李隽颖(1982-)，男，湖北云梦人，博士研究生，主要研究方向为振动信号特征提取和特征选择(anglee2007@126.com)；楼晓俊(1984-)，男，浙江杭州人，博士研究生，主要研究方向为无线传感器网络.

半径。球面上的点对球体的确定起关键作用,称其为支持向量。文献[9]阐述了基于序贯最小优化(sequential minimal optimization,SMO)算法的超球支持向量机单类样本训练算法的实现。

2003 年,朱美琳等人[10]将超球支持向量机扩展到多分类情况,在多分类情况下为每类样本建立一个超球体支持向量机模型,在空间中形成像肥皂泡一样的分类结构,判决时,空间中的点到哪个球最近,则属于哪个类。2009 年,徐图等人[11]完善了超球体支持向量机的多分类理论,并提出了基于 SMO 的多超球支持向量机训练算法,并用数据证实了超球体支持向量机解决多分类问题训练速度明显快于一对多方法[2]、一对一方法[3],验证了超球支持向量机是一种非常适合于多分类的分类器。

对于多分类问题,本身求解就比较复杂,当特征向量的维数过大时,不仅导致较大的计算量,并且也会带来分类识别率的下降。2011 年,Pineda - Bautista 等人[12]提出了基于特征选择的多分类器设计框架,本文在该框架下,使用基于排挤的小生境遗传算法实现特征向量的选取,为不同的目标类别寻找最优的特征子集。由于不同特征对不同目标识别的贡献程度不一样,这种方式更有利于提升超球支持向量机的分类性能。

1 基于特征选择的 HSSVM 分类方法

随着模式识别研究的深入,研究对象具有多类别、高特征维数等越来越复杂的特点。大量高维数据对象的特征空间含有许多冗余特征甚至噪声特征,这些特征一方面可能降低分类的精度,另一方面会大大增加分类算法的时间及空间复杂度。因此,在对高维数据进行分类时,运用特征选择算法找到具有较好可分性的特征子空间,同时合适的参数设置也会提升分类器性能。本文使用基于排挤的小生境遗传算法实现特征向量的选取,为不同的目标类别寻找最优的特征子集。对于多分类问题,由于不同特征对不同识别类别的贡献程度不一样,这种方式更有利于提升超球体多类支持向量机的分类性能。

如图 1、2 所示,以三分类为例, 基于特征选择的 HSSVM 的分类模型建立过程和分类判决过程。对于样本 W_1,特征选择的目的是找到一个最优特征向量的子集,使得训练出的 HSSVM1 能最大程度地把 W_1 样本包含在球面内,把 W_2、W_3 的样本排除在球面以外,对另外两类样本进行特征选择也是按照该原则。

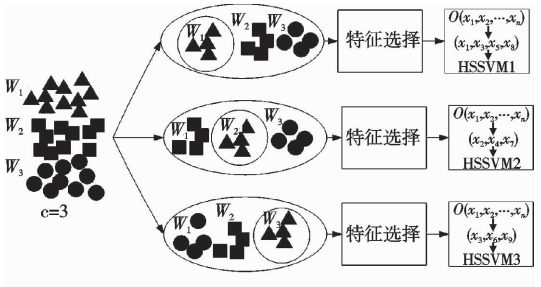

图1 基于特征选择的HSSVM分类模型

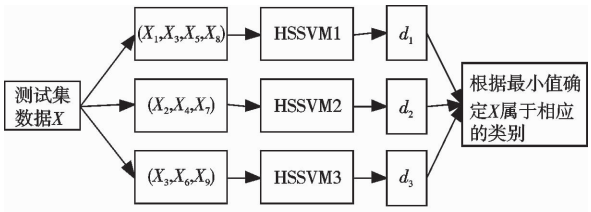

图2 基于特征选择的HSSVM分类判决过程

测试阶段,计算测试点 x 映射到高维空间后到每个球心的距离,将距离除以相应的球半径,作为该点与每个球之间的度量 $D=\{d_1,d_2,d_3\}$,即

$$d_i=\frac{\|\Phi_i(x)-a_i\|}{R_i} \tag{6}$$

其中:$\Phi_i(x)$ 是 x 映射到球面 i 核函数空间的坐标;a_i 是球面 i 的球心;R_i 是球面 i 的半径。若 $d_i=\min(D)$,则 x 属于第 i 类。

对于判别方法,一般通过找出测试点到各球心的距离最短的超球来判断测试点属于哪一类,在遇到半径大小相差较大的球体之间,点到球心距离相差不大的情况下,这种方法存在缺陷。本文所用判定方法是将距离和球的半径联系起来,对于半径相差较大的球体之间起到平衡作用,判决更为合理。

2 基于小生境遗传算法的特征选择方法

2.1 小生境遗传算法介绍

遗传算法[13]模拟了自然选择和遗传中发生的复制、交叉和变异现象,从任一初始群体出发,通过随机选择、交叉和变异操作,产生一群更适应环境的个体,使群体进化到搜索空间中越来越好的区域,这样一代一代地不断繁衍进化,最后收敛到一群最适应环境的个体,求得问题的最优解。普通的遗传算法对于一个优化问题只能发现一个最优解,很大可能是局部最优解。群体中个别个体大量集中容易导致传统遗传算法的早熟收敛,虽然交叉和变异可以产生新的个体,但是不能改变个体大量积聚的状况,搜索空间解不能有效跳出当前解空间,造成进化停滞不前,搜索空间解陷入局部寻优的困境。

遗传算法用于特征选择的伪代码算法流程:a)初始化当前进化次数;b)初始化种群个体;c)计算初始种群每个个体的适应度;d)选择父代个体;e)用选择的父代个体繁殖下一代个体;f)计算下一代个体的适应度;g)结合两代个体选择最优个体;h)进入下一轮进化。具体如下所示:

```
a) count = 0;
b) X_count = {x_0, x_1, ···, x_n};
c) F_t1 = Eval(X_count);
while(not done)
{
d) P_count = select(X_count, F_t1);
e) O_count = procreate(P_count);
f) F_t2 = Eval(X_count);
g) X_count = select_surivors(X_count, F_t1, O_count, F_t2);
h) count = count + 1;
}
```

最初为了防止遗传算法早熟收敛,引入了基于预选择机制的小生境技术,增加种群中的差异度。小生境技术源于遗传算法,在遗传算法中,发现多模问题中多个最优解的能力被称为小生境技术,该技术保证了群体多样性在进化过程中维持相对的稳定,并且改善了传统遗传算法的全局寻优能力。

常见的有基于适应度共享的小生境技术[14,15]、基于排挤的小生境技术和基于预选择的小生境技术。适应度共享的小生境技术定义了共享函数来调整群体中各个个体的适应度,从而在以后的群体进化过程中,算法可以依据调整以后的新适应度进行选择运算,以维护群体的多样性,创造小生境的进化环境。该算法最主要的缺陷是需要先验知识设定小生境半径。

根据种群多样性的变化,本文所采用的基于排挤的改进小

生境遗传算法对保持种群多样性与全局寻优性能起到积极作用。方法的基本思想是:由群体中随机选择若干个体组成排挤成员,依据新产生的个体与排挤成员的相似性来排挤掉一些与排挤成员相似的个体。随着排挤过程的进行,群体中个体逐渐被分类,从而形成各个小的生存环境,并维持了种群的多样性。

2.2　ICNGA 基本思想

2.2.1　二进制个体编码及适应度函数

若 n 为特征空间的维数,特征选择的二进制编码可用 $X=\{x_1, x_2, \cdots, x_n\}$, $x_i \in \{0,1\}$ $(i=1, 2, \cdots, n)$ 表示。如果 X 的第 i 位为 1,则此特征被选中;如果为 0,则此特征未被选中。

在超球支持向量机的分类问题中,期望在分类效果最好的情况下,选择最少的特征维数。

假设适应度函数为

$$F(e,f)=\frac{n_s-e}{n_s}-\frac{f}{n_f}\cdot\alpha \tag{7}$$

其中:e 为训练样本出现错误分类的个数;n_s 为训练样本总数;f 为选择出的特征子空间的维数;n_f 为特征空间的维数;α 为待确定的参数。对于 e 要优先保证分类正确率,再保证特征子空间的维数更少,因此式(7)需满足

$$F(e-1,n_f)>F(e,f)>F(e+1,1) \tag{8}$$

其中,$1\leqslant f\leqslant n_f$。

可以推导出 $\alpha=\frac{1}{n_s}$是一个较好的参数。

适应度函数可以表示为

$$F(e,f)=\frac{n_s-e}{n_s}-\frac{f}{n_s\cdot n_f} \tag{9}$$

2.2.2　初始化种群

为了保证初始化种群的多样性,增强全局寻优的能力,采用了排挤式筛选算法。针对 2.1 节中步骤 a)b)作修改。随机初始化两个种群 $X_0=\{x_0,x_1,\cdots,x_n\}$, $X_0'=\{x_0',x_1',\cdots,x_n'\}$,并计算出两个种群的个体适应度 $F_0=\mathrm{eval}(X_0)$, $F_0'=\mathrm{eval}(X_0')$;将 X_0'中的个体 $X_0'(i)$, $\{i=0,1,\cdots,n\}$, 在 X_0 中寻找与其海明距离最近的个体 $X_0(j)$, $\{j=0,1,\cdots,n\}$。若 $F_0(j)<F_0'(i)$,则个体 $X_0'(i)$取代 $X_0(j)$, $F_0'(i)$取代 $F_0(j)$;否则不改变。最终输出更新之后的 X_0 和 F_0。

2.2.3　遗传操作算子的设计

这部分涉及 2.1 节中步骤 d) ~ f)。根据调整之后的新种群 X_{count}按适应度 F_{t1}进行轮盘赌法,并按一定的比率选择个体进行交叉和变异操作,生成子代个体 O_{count},并计算出相应的个体适应度 F_{t2}。将 O_{count}中的个体 $O_{\mathrm{count}}(i)$, $\{i=0,1,\cdots,n\}$,在 X_{count}中寻找与其海明距离最近的个体 $X_{\mathrm{count}}(j)$, $\{j=0,1,\cdots,n\}$。若 $F_{t2}(i)>F_{t1}(j)$,则个体 $O_{\mathrm{count}}(i)$取代 $X_0(j)$, $F_{t2}(i)$取代 $F_{t1}(j)$;否则不改变。最终输出更新之后的 X_{count}和 F_{t1}。

2.3　ICNGA 实现步骤

a)选择二进制编码规则、训练样本和测试样本。

b)初始化参数,即遗传代数、种群大小、交叉概率、变异概率。

c)随机生成两个种群,对每个个体进行解码变换,计算个体适应度,个体适应度通过本类训练样本训练出一个超球支持向量机,通过所有训练样本来测试有多少样本会被错误分类,获取式(9)中的 e;另外,n_s 是训练样本总数,f 是选择出的特征子空间的维数,n_f 是特征空间的维数,依据式(9)计算出个体适应度。

d)根据 2.2.2 节,将上述两个种群通过排挤的方式合并成一个种群。

e)根据调整之后的适应度按轮盘赌法机制,并按一定的比率选择个体进行交叉、变异操作产生子代种群。

f)对每个个体进行解码变换,按 2.1 节中步骤 e)计算个体适应度,根据 2.2.3 节合并父代和子代种群。

g)收敛性判定,如果满足收敛条件,或已经进化了规定的代数,则结束进化过程,保存获取的最优参数;否则,返回 e)。

h)按以上规则直至所有类的特征选择和超球支持向量机训练完毕,按照图 2 流程和式(6)对测试数据进行分类实验,获取最终分类结果。

3　数值实验

根据上述基本思想,设计了一个个体排挤规则,不需要先验的知识就可以动态地调整群体的规模,建立小生境环境,同时保证获得更多的局部最优解,增大寻找到全局最优解的概率。下面是利用 UCI 数据库中的 Wine 和 Arrhythmia 数据集进行的数值实验。表 1 为选用的数据集特性。

表 1　数据特性

数据集名称	样本总数	特征维数	类别数	训练样本数	测试样本数	数据缺失
Wine	178	13	3	90	88	否
Arrhythmia	452	279	3	235	217	是

在计算个体适应度时,对单类训练样本用基于 SMO 的超球支持向量机的训练方法[9],并用其他类的训练样本来验证训练出的超球支持向量机的分类性能来验证该个体对应的选择出的特征向量是否最优。在训练超球支持向量机的过程中,采用了径向基核函数,表达式为

$$K(x,y)=\exp(-\|x-y\|^2/2\sigma^2) \tag{10}$$

设 $C=0.8$, $\sigma^2=100$, $e=0.02$。针对 Wine 数据集和 Arrhythmia 数据集,分别用特征全选的分类方式,用常规遗传算法、常规小生境遗传算法的特征选择分类方式和本文提出的小生境遗传算法特征选择分类方式进行测试。其中遗传算法和小生境遗传算法中的参数设置为:种群规模为 30,最大进化代数为 50,交叉概率为 0.6,变异概率为 0.1。对于 Arrhythmia 数据集,其样本某些维的特征数据存在缺失,对缺失数据作如下处理:

a)若缺失数据为训练样本中某一维的特征数据,则取该类训练样本的所有该维的特征数据的均值为缺失数据值。

b)若缺失数据为测试样本中某一维的特征数据,则取所有训练样本的所有该维的特征数据的均值为缺失数据值。

图 3、4 分别为 Wine 和 Arrhythmia 数据集对应三个类的样本进行特征优化的种群适应度进化曲线,分别对不作特征选择(即特征全选)、GA 进行特征选择、NGA[16] 进行特征选择和 ICNGA 进行特征选择的个体适应度进化情况作了比较。

从图 3、4 可以看出,经过特征选择之后的个体适应度高于不作特征选择的个体适应度;GA 的收敛速度比 NGA 和 ICNGA 要快,但 ICNGA 的寻优能力最高,NGA 次之。证实了 ICNGA 的全局寻优能力要好于 GA,略优于 NGA。

表 2 所示分别是不作特征选择的分类结果、GA 进行多次优化的平均分类结果和 NGA 进行多次优化的平均分类结果。

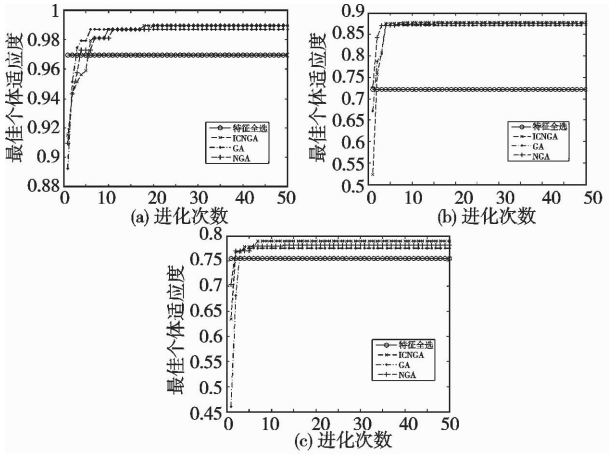

图3 Wine数据集的三类训练样本的种群适应度进化曲线

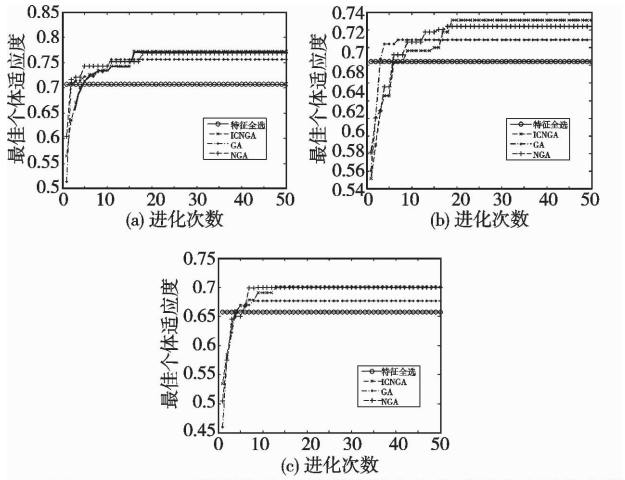

图4 Arrhythmia数据集的三类训练样本的种群适应度进化曲线

表2 分类识别正确率

数据集	特征全选/%	GA/%	NGA/%	ICNGA/%
Wine	82.6	85.6	86.3	86.8
Arrhythmia	64.5	66.8	70.1	71.5

从表2可以看出,经过特征选择之后的分类效果好于不作特征选择的分类效果;ICNGA优化之后的分类效果要好于GA和NGA优化之后的分类效果;在维数高的Arrhythmia数据集中相比维数较低的Wine数据集,ICNGA相对GA和NGA提升的分类性能更为明显。

4 结束语

提出了将NGA和HSSVM结合应用于多分类问题优化。构造了一种基于排挤的小生境算法(ICNGA)。该算法就数据样本特征进行优化,为不同的目标类别寻找最优的特征子集,提高了分类器性能。构造的ICNGA算法具有较强的全局优化能力。尤其在分类数据类别较多、特征维数较高的情况时,相比于不作特征选择、GA和常规NGA能获得更好的分类效果。

参考文献:

[1] VAPNIK V N. The nature of statistical learning theory [M]. New York: Springer-Verlag, 1995.

[2] VAPNIK V N. Statistical learning theory [M]. New York: Wiley, 1998.

[3] KREBEL U H G. Pairwise classification and support vector machines [M]//Advances in Kernel Methods. Cambridge, MA: MIT Press, 1999:255-268.

[4] GUO Guo-dong, LI S Z, CHAN K. Face recognition by support vector machines[C]//Proc of the 4th International Conferences on Automatic Face and Gesture Recognition. 2000: 196-201.

[5] SCHWENKER F, PALM G. Tree-structured support vector machines for multi-class pattern recognition[C]//Proc of the 2nd International Workshop on Multiple Classifier Systems. London: Springer-Verlag, 2001: 409-417.

[6] BREDENSTEINER E J, BENNETT K P. Multicategory classification by support vector machine [J]. Computational Optimization and Applications, 1999, 12(1-3): 53-79.

[7] SCHOLKOPF B, BURGES C, VAPNIK V. Extracting support data for a given task[C]//Proc of the 1st International Conference on Knowledge Discovery & Data Mining. Menlo Park, CA: AAAI Press, 1995:252-257.

[8] TAX D M J, DUIN R P W. Data domain description by support vectors[C]//Proc of European Symposium on Artificial Neural Networks. 1999:251-256.

[9] 徐图,罗瑜,何大可. 超球体单类支持向量机的SMO训练算法[J]. 计算机科学, 2008,35(6):178-180.

[10] 朱美琳,刘向东,陈世福. 用球结构的支持向量机解决多分类问题[J]. 南京大学学报:自然科学版, 2003,39(2):153-158.

[11] 徐图,何大可. 超球体多类支持向量机理论[J]. 控制理论与应用,2009,26(11):1293-1297.

[12] PINEDA-BAUTISTA B B, CARRASCO-OCHOA J A, MARTINEZ-TRINIDAD J F. General framework for class-specific feature selection [J]. Expert Systems with Applications, 2011, 8(38): 10018-10024.

[13] 史忠植. 知识发现[M]. 2版. 北京:清华大学出版社,2011:346-374.

[14] 乔佩利,郑林,马丽丽. 一种小生境遗传算法研究[J]. 哈尔滨理工大学学报, 2011,16(1):90-93.

[15] 朱宁,冯志刚,王祁. 基于小生境遗传算法的支持向量机分类器参数优化[J]. 南京理工大学学报:自然科学版, 2009,33(1): 16-20.

[16] 黄浩,宋瀚涛,陆玉昌. 基于小生境遗传算法的贝叶斯网络结构学习算法研究[J]. 计算机应用研究, 2007, 24(4):100-103.

下期要目

- 个性化推荐系统隐私保护策略研究进展
- 基于视觉的移动机器人可通行区域识别研究综述
- 中文文本情感分析研究综述
- 软件移植理论与技术研究
- 英语语篇结构分析研究综述
- 一种新的混合粒子群算法求解置换流水车间调度问题
- 基于改进蚁群算法的车辆路径优化问题研究
- 自适应阶段变异量子粒子群优化算法研究
- 一种结合人工蜂群和K-均值的混合聚类算法
- 基于遗传算法的数据交换问题求解方法
- 全信息差异进化粒子群优化算法
- 一种新的基于logistic混沌映像的自适应混沌蚁群优化算法求解动态车辆路径问题
- 基于多链扩展编码方案的量子遗传算法
- 基于组织进化粒子群优化的测试用例自动生成
- 基于选择和变异机制的蛙跳FCM算法
- 求解无等待流水车间调度优化的一种启发式算法