

基于多群协同人工鱼群算法的分类规则提取算法*

戴上平¹, 姬盈利¹, 王 华², 金 鹏¹

(1. 华中师范大学 计算机科学系, 武汉 430079; 2. 武汉大学 资源与环境科学学院, 武汉 430079)

摘 要: 在对基本人工鱼群算法原理分析的基础上, 提出了一种多群协同人工鱼群算法用于实现对连续空间变量的分类规则提取问题。定义了基于规则支持度与置信度的规则评价函数, 构造了人工鱼在规则提取应用中的特定编码及相关概念的计算公式, 给出了该算法的具体实现步骤, 并用 VC++ 软件编程实现。最后对 Iris 和 Wine 数据集进行测试实验, 并与单群体鱼群算法及多种群微粒群算法进行比较。仿真结果表明, 该算法能够快速提取分类精度较高的分类规则, 因此利用该算法解决连续变量分类规则提取的相关问题是可行且有效的。

关键词: 多群体人工鱼群; 分类规则; 协同; 单群体人工鱼群算法; 多种群微粒群算法

中图分类号: TP301.6 **文献标志码:** A **文章编号:** 1001-3695(2012)05-1676-04

doi:10.3969/j.issn.1001-3695.2012.05.020

Extracting classification rules based on multi artificial fish swarm cooperation algorithm

DAI Shang-ping¹, JI Ying-li¹, WANG Hua², JIN Peng¹

(1. Dept. of Computer Science, Central China Normal University, Wuhan 430079, China; 2. School of Resource & Environment Science, Wuhan University, Wuhan 430079, China)

Abstract: This paper proposed the multi artificial fish swarm cooperation algorithm (MAFWA) based on the work principle of basic artificial fish swarm to extract the classification rules of continuous variable space. It defined a function to evaluate the rules based on the support and confidence, designed the code of artificial fish, defined the formula to calculate some key parameters for its application in extracting classification rules and provided with the detailed step of algorithm. And then it made use of programming in the VC++ 6.0. Finally, it made an experiment on Iris and Wine data sets to test the algorithm, further more it compared the algorithm of MAFWA with the single artificial fish swarm and multi particle swarm algorithm by using the same data sets of Iris and Wine. The experimental result shows that this algorithm can extract the classification rules with high precision quickly. So the algorithm of MAFWA is available and efficient to resolve the problems related to extracting classification rules of continuous variable.

Key words: multi artificial fish swarm algorithm; classification rule; cooperation; single artificial fish swarm algorithm; multi particle swarm algorithm

人工鱼群算法是由李晓磊等人^[1]于2002年提出的一种全局群智能优化算法^[1]。该算法在研究鱼群活动的基础之上, 通过构造人工鱼、模仿鱼群底层的觅食、聚群、追尾等行为来实现寻优^[2-4]。算法能够较好地获取全局极值, 并具有对初值参数选择要求不高、对启发式函数不敏感、简单易实现、快速性等优点^[5-8], 因此被广泛应用于解决函数优化^[9]、约束优化^[10]、参数估计^[11]、模式识别^[12]、分类预测^[13]等相关问题。将鱼群算法用于解决分类规则提取问题的研究还很少见, 有相关文献介绍了此算法在分类规则挖掘中的应用, 但是由于其编码的设计只能对离散数据进行分类规则提取^[14], 因此本文尝试利用人工鱼群算法对连续空间变量进行分类规则提取。同时受多微粒群协同进化算法思想的^[15]启发, 本文提出了基于多群协同人工鱼群(MAFWA)进行分类规则提取的算法。与传统鱼群算法采用单鱼群提取多分类规则方法相比, 本文利用多种群人工鱼群中的每一鱼群对应于一条规则的提取, 其收敛速度更

快、求解精度更高。并且由于多种群协同人工鱼群在数据以及任务分割上的简便性, 使其在将来算法的并行性实现上更为简单。

1 基本人工鱼群算法

鱼类通过觅食来满足自身成长过程中营养物质的需要。自然现象表明, 水域中营养物质多的地方一般也就是大量鱼类的聚集地^[1]。对鱼类生存影响比较大的行为有觅食、聚群、追尾行为, 人工鱼群算法就是基于鱼类觅食的这一特性, 通过构造人工鱼模仿鱼类的这些行为来实现寻优的。

1.1 算法相关定义

假定人工鱼当前的状态表示形式为 $x = (x_1, x_2, x_3, \dots, x_n)$, ρ 表示拥挤度因子, 人工鱼当前所处环境的食物浓度为 $y = f(x)$, $step$ 表示每次移动的步长, $visual$ 表示人工鱼的感知范围。为了避免算法在后期处于平坦区域时容易陷入局部极

收稿日期: 2011-10-30; 修回日期: 2011-11-30 基金项目: 国家“863”计划资助项目(2011AA120304)

作者简介: 戴上平(1962-), 男, 湖北武汉人, 副教授, 博士, 主要研究方向为人工智能、数据挖掘(lucky201199@163.com); 姬盈利(1986-), 女, 河南洛阳人, 硕士研究生, 主要研究方向为数据挖掘、人工智能; 王华(1986-), 男, 河南信阳人, 博士研究生, 主要研究方向为土地信息系统研究开发、空间数据挖掘; 金鹏(1988-), 男, 湖北新洲人, 硕士研究生, 主要研究方向为自然语言处理、人工智能。

值、收敛速度慢的缺陷,采用自适应的步长和视野,在算法初期或者食物浓度随着迭代次数不发生显著变化陷入局部极值时采用较大的步长和视野,从而更好地获取全局最优解;在算法后期采用较小的步长和视野,从而有利于获得较高的求解精度。

1.2 人工鱼个体的行为描述

1.2.1 觅食行为

觅食行为是鱼类生存的本能行为。鱼类在水中自由游动,并通过感官向食物浓度大的地方游去。设定人工鱼当前状态为 x_i ,在其邻域($d(i,j) < \text{visual}$)内随机选择另外一个状态 x_j ,代入目标函数计算所在点的食物浓度分别为 y_i, y_j 。如果 $y_i < y_j$,则表明 x_j 处相对 x_i 处有较大的食物浓度,则人工鱼向 x_j 处前进一步;否则就再选择一个新的状态 x_j ,重新比较 x_j 与人工鱼当前位置的食物浓度大小。如果试探了 trynumber 次仍然找不到食物浓度比较大的点,就随机移动一步。

1.2.2 聚群行为

基于动物的群居特性有利于共同防御敌害或者集体觅食。鱼类通常成群活动,在聚群游动的过程中既要保证鱼群周围有较大的食物浓度,同时也要注意鱼群之间不要过分拥挤。设定人工鱼当前状态为 x_i ,在其邻域($d(i,j) < \text{visual}$)内搜索中心位置 x_c 及伙伴数目 n_f ,代入目标函数计算所在点的食物浓度分别为 y_i, y_c 。如果 $y_c/n_f > \alpha \times y_i$,则表明伙伴中心有较多的食物,并且不太拥挤,则向伙伴中心位置随机移动一个步长;否则执行觅食行为。

1.2.3 追尾行为

当鱼群中的某个个体找到食物时,通过简单的通信传递信息,其他伙伴会尾随其后,朝着鱼群中距离自己较近的最优个体靠拢。设定人工鱼当前状态为 x_i ,在其邻域($d(i,j) < \text{visual}$)内搜索食物浓度 $y_{\max}$ 最大的人工鱼 $x_{\max}$ 。如果 $y_{\max}/n_f > \alpha \times y_i$,说明 $x_{\max}$ 处有较多的食物,并且不太拥挤,则向 $x_{\max}$ 方向随机移动一个步长;否则执行觅食行为。

1.2.4 行为评价

人工鱼通过迭代模仿鱼类的行为,并且通过自身活动对同伴状态和环境状态产生影响,在每次迭代后判断比较自身当前的状态、当前环境的状态与同伴的状态,选择模仿实现鱼的觅食、聚群、追尾、随机行为中能使食物浓度达到最大的一种行为来执行。这样在不断的比较判断更新自身状态的过程中,使得人工鱼的下一状态最优,人工鱼个体通过这样的方法局部寻优,最终使全局最优值在群体中突现出来。

2 多种群协同人工鱼群分类规则提取算法

分类的目的是通过对目标数据集合的机器学习而得出一个分类函数或分类模型(即分类器),然后利用分类器将数据对象映射到一个给定的类别中。对分类规则的提取,其实质就是构造一个由规则集合组成的可用于对数据进行特征区分的分类器。一般规则提取算法主要包括三个部分:数据采样→规则提取→规则测试。数据采样部分通过多种采样方法进行数据集合的划分:一部分称之为训练数据集合,用于规则提取训练;一部分称之为测试数据集合,用于规则精度测试。训练数据集合部分通过分析由属性描述的数据集合;规则提取部分通

过算法的演绎迭代反映数据集合特性的属性特征,即规则逐步提取出来加入到规则集合中;规则测试部分使用测试数据集合进行规则集合的准确度和精度评估。

多种群协同人工鱼群分类规则提取算法的特点是通过将数据与任务进行分割,每一个鱼群对应负责某一类数据特征的提取任务,多个鱼群并行计算,鱼群之间无须进行复杂的通信,在满足一定规则精度或者达到预先设定的迭代次数时,再将各鱼群提取的规则进行综合,便形成完整的置信度较高的用于分类预测的规则集合,如图1所示。为实现多种群协同人工鱼群分类规则提取算法,需明确分类规则表达形式、分类规则评价、隶属冲突解决、人工鱼构造四类问题。

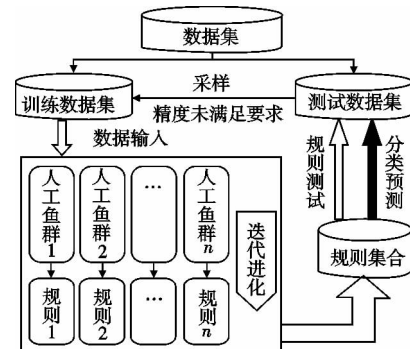


图1 多种群协同人工鱼群分类器示意图

2.1 分类规则表达形式

一般分类的规则表示形式为:if (A_i) then (C_i) with RP (A_i, C_i)。式中:规则前件 A_i 代表数据集合的特征属性集合,规则后件 C_i 代表数据集合类别;RP (A_i, C_i) 代表隶属规则的优先级,用于解决某个数据个体符合多项判别规则的问题。当某个实例的特征属性集合与规则前件匹配,并且用该规则分类得出的实例类别与该实例所属的类别一致时,就说明这个实例被正确分类^[16]。

2.2 分类规则评价

要提取到最优的规则,就需在训练过程中确定合适的适应度函数对规则进行评估。规则的支持度(support)和规则的可信度(confidence)是衡量规则好坏的两个重要指标,因此在确定适应度函数之前引入这两个指标的相关定义。

定义1 支持度。对于规则 if (A_i) then (C_i),如果某种群中有% s 的个体满足规则,即包含 $A_i \cup C_i$,则称该规则的支持度为% s ,记为 $\text{support}(A_i)$,即

$$\text{support}(A_i \Rightarrow C_i) = p(A_i \cup C_i) \quad (1)$$

定义2 置信度。若规则 if (A_i) then (C_i) 的支持度为 $\text{support}(A_i)$,则规则的信任度为 $\text{support}(A_i \cup C_i)/\text{support}(A_i)$,记为

$$\text{confidence}(A_i \Rightarrow C_i) = p(C_i | A_i) \quad (2)$$

定义1说明规则的支持度及规则的普遍性反映了正确分类的个体数目占种群中个体总数的比例。规则的支持度越高,表明该规则能将实例进行正确分类的数目越多。定义2说明规则的信任度即规则的准确度反映一条规则能将一个个体正确分类的概率。规则的信任度越高,表明个体用该规则能够正确分类的概率越大。因此确定适应度函数为

$$F(x) = (a \times \text{support}(A_i \Rightarrow C_i)) + ((1-a) \times \text{confidence}(A_i \Rightarrow C_i)) \quad (3)$$

式(3)通过支持度和置信度来综合评估规则的质量。式中 a 为支持度系数, 值域为 $(0, 1)$ 。通过该系数保证生成的规则即能将大部分的样本正确分类, 同时也有较高的概率能将单个样本正确分类, 从而保证生成的规则向最优的方向发展。

2.3 隶属冲突解决方案

当一个样本 X 特征同时满足多个规则时, 由于该样本只能隶属一个规则的结论, 这样就会产生该样本的隶属冲突。利用式(4)计算该样本与规则覆盖样本集的距离 $d(x)$, 选择距离最小的规则的结论作为该样本的结论。

$$d(x) = \sqrt{\sum_n (X_n - \bar{X}_n)^2} \quad (4)$$

其中: $\bar{X}_n$ 为平均样本。

2.4 人工鱼构造

为了将人工鱼用于连续空间变量的数据分类规则提取问题中, 需要在基本人工鱼算法的基础上对人工鱼编码、人工鱼距离、人工鱼的邻居及聚类中心进行相关的定义。

2.4.1 人工鱼编码

本算法中每条人工鱼代表一条候选规则, 在算法实现中, 对人工鱼的编码需要能够直观地表达规则, 因此用 $Afish(X_1, X_2, \dots, X_n, C)$ 这种编码形式来代表一条人工鱼。其中前 n 维的具体编码见表1, 每一维均分别定义其特征上下界; 第 $n+1$ 维 C 代表类别属性。将其解码为规则表现形式: if $X_{1_min} < X_1 < X_{1_max}$ and $X_{2_min} < X_2 < X_{2_max}$ and $\dots$ and $X_{n_min} < X_n < X_{n_max}$ then C 。

表1 人工鱼编码

特征属性	X_1	X_2	$\dots$	X_n
上界	X_{1_max}	X_{2_max}	$\dots$	X_{n_max}
下界	X_{1_min}	X_{2_min}	$\dots$	X_{n_min}

2.4.2 人工鱼的相关概念

在算法的迭代进化过程中, 人工鱼距离、邻域、聚群中心位置、食物浓度是几类非常重要的概念。本文定义了这四类概念在分类规则提取中的具体实现方式。

a) 人工鱼距离。它表示两个规则之间的差异性, 本文定义了基于欧几里德距离的人工鱼距离计算式。例如人工鱼 $Afish_i$ 与 $Afish_j$ 之间的距离可通过下式来计算:

$$d(i, j) = \sqrt{\sum_{k=1}^n (X_{ik_max} - X_{jk_max})^2 + (X_{ik_min} - X_{jk_min})^2} \quad (5)$$

其中: n 代表特征维数; X_{ik_max} 代表人工鱼 i 的第 k 个属性的上界; X_{ik_min} 代表人工鱼 i 的第 k 个属性的下界。

b) 邻域。在人工鱼距离的基础上定义人工鱼 $Afish_i$ 的邻域:

$$n(i, \varphi) = \{i' | d(i, i') < \varphi, G \ni i'\} \quad (6)$$

其中: i' 代表邻域内其他个体; φ 代表邻域距离; G 代表整个鱼群。

c) 中心位置。一个人工鱼聚群中的中心位置, 在算法中分别取所有人工鱼的每一维特征值上、下界的平均值。

$$c(X_1, X_2, \dots, X_n) = \left(\frac{\sum_{k=1}^n X_{k1}}{n}, \frac{\sum_{k=1}^n X_{k2}}{n}, \dots, \frac{\sum_{k=1}^n X_{kn}}{n} \right) \quad (7)$$

式中: n 代表聚群中人工鱼个体数。

d) 食物浓度。它表示规则质量, 通过式(3)进行计算。

2.5 整体算法描述

在上述定义的基础上, 本算法的整体实现流程(图2)为:

a) 通过数据采样将数据集分为训练数据集和测试数据集。

b) 设置鱼群欲提取规则的类别及步长、可视范围等参数, 根据2.4.1节的人工鱼编码方式进行种群的初始化。

c) 对鱼群按式(5)~(7)来进行觅食、追尾、聚群等行为。

d) 根据式(3)计算每一人工鱼的适应值, 对规则质量进行评估, 将最优个体放入公告板。

e) 判断算法是否达到最大迭代次数或者连续几代公告板未发生变化, 则认为达到算法终止条件执行f), 否则执行c)。

f) 将生成的规则加入到规则集合中。

重复执行步骤a)~f), 直到找到较优规则。最后用测试数据集检验规则集合的精度, 如果不满足要求则重新进行步骤a)。

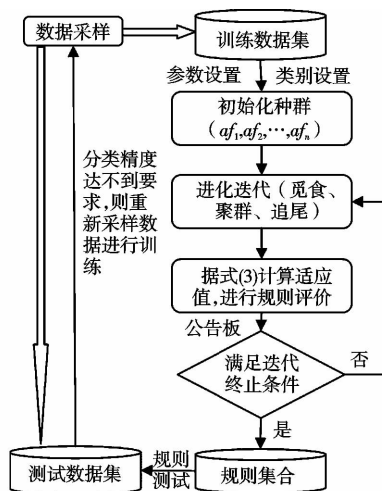


图2 多种群协同人工鱼群分类规则提取算法流程

3 实验仿真及分析

3.1 实验数据集

为验证算法的可行性, 选取 UCI 知识库中的 Wine、Iris 数据集来进行实验。植物数据集 Iris 包含 150 个样本, 每个样本包含 4 个特征属性, 即花瓣长度、花瓣宽度、花萼长度、花萼宽度, 根据 4 个特征属性将其划分为 3 个类 (setosa、versicolor、virginica)。Wine 数据集包含 178 个样本, 每个样本包含 13 个特征属性、3 个类别, 如表2所示。

表2 UCI 数据集

数据集	样例数(按分类分布)	特征数目	类别数目
Iris	150 (50 + 50 + 50)	4	3
Wine	178 (59 + 71 + 48)	13	3

3.2 实验结果与分析

本文采用的是 5 折交叉验证法进行实验, 用 $n\%5$ 的策略把数据分成 5 份, 其中 n 为数据的记录号, 即将行号除以 5 所得余数相等的数据归为一个子集, 设为 $D_1, D_2, \dots, D_5$ 。训练和测试轮流进行 5 次, 在每一次训练都选用不同的数据子集作为测试集, 其余的 4 个子集都用于训练集, 取 5 次的平均值作为最终的实验结果。

本算法的参数设置如下: 最大迭代次数 = 50, 单群人工鱼个数为 30, 可视范围 = 2, 步长 = 0.5, 最大尝试次数 = 30, 支持度系数为 0.5。

为比较算法的改进效果,本文还实现了基于单鱼群(single artificial fish swarm algorithm,SAFWA)的分类规则提取算法,其主要区别在于将人工鱼的编码扩展为多类别规则表现形式,主要参数与本文算法保持一致。计算结果如表3所示。

表3 MAFWA与SAFWA分类结果对比

算法/数据集	分类精度/%		计算时间/ms	
	Iris	Wine	Iris	Wine
MAFWA	94.3	91.58	850	1 920
SAFWA	91.75	87.9	1 160	2 630

由结果可以看出,无论是分类精度还是计算时间,MAFWA均优于SAFWA,其主要原因是MAFWA中的每一鱼群只用于提取一个特征类别的规则,相对于同时提取所有类别特征的SAFWA,MAFWA的人工鱼编码简单很多,在计算的复杂性与较优解的搜索上具有优势。

本文还给出了MAFWA对Iris数据集提取规则的结果:

if $4.22 \leq x_1 \leq 6.16$ and $2.44 \leq x_2 \leq 4.48$ and $1.09 \leq x_3 \leq 2.21$ and $0.13 \leq x_4 \leq 0.79$ then class = setosa;

if $4.86 \leq x_1 \leq 6.8$ and $2.06 \leq x_2 \leq 3.51$ and $2.75 \leq x_3 \leq 5.36$ and $1.13 \leq x_4 \leq 1.84$ then class = versicolor;

if $4.83 \leq x_1 \leq 7.74$ and $2.07 \leq x_2 \leq 4.38$ and $4.62 \leq x_3 \leq 7.13$ and $1.18 \leq x_4 \leq 2.54$ then class = virginica.

为更好地了解算法的收敛性及寻优性能,本文跟踪了MAFWA算法对Iris数据集的适应值与规则分类精度的变化趋势,如图3所示。

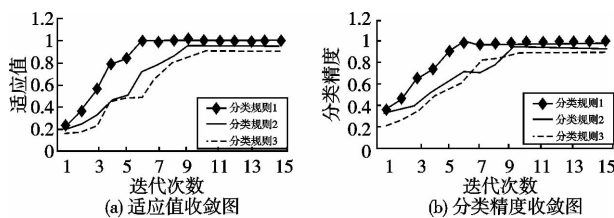


图3 Iris数据集的适应值与分类精度收敛图

由图3可以看出,每一类都分别提取出一条规则,其中规则1在第8代开始收敛,对类别setosa的分类精度达到100%;规则2在第9代开始收敛,分类精度达到约96%;规则3则在11代开始收敛,分类精度约为93%。

同时本文还将分类结果与文献[17]中的多群体微粒群算法(MPSO)进行了比较,如表4所示。

表4 MAFWA与MPSO分类结果对比

算法/数据集	分类精度/%		规则数目/个	
	Iris	Wine	Iris	Wine
MAFWA	94.3	91.58	3	4
MPSO	94.53	93.1	5	4

由表4可以看出,在分类精度方面,在Iris数据集上两类算法分类精度接近;在Wine数据集上,MAFWA的分类精度要稍低,说明MAFWA在对多维特征数据集的搜索最优解的能力要稍弱于MPSO算法。在规则数目方面,两类算法对Wine数据集均提取了四条规则,MAFWA对Iris数据集提取了三条规则,而MPSO提取了五条规则,说明MAFWA在低维数据集的学习能力不逊于MPSO。

4 结束语

通过对Wine和Iris数据集的仿真实验说明利用MAFWA

方法可以快速得到规则简单、精度较高的分类结果,并且相比单群体鱼群算法,不论是在计算时间和规则精度上都具有明显的优势;与MPSO相比较,除在多维数据Wine的规则提取精度上有所逊色之外,对低维数据Iris的学习能力较优于MPSO算法。因此该算法在对连续变量空间分类规则提取问题中是可行且有效的。下一步着重研究如何提高该算法在高维连续数据空间内的分类能力。

参考文献:

- [1] 李晓磊,邵之江,钱积新.一种基于动物自治体的寻优模式:鱼群算法[J].系统工程理论与实践,2002,22(11):32-38.
- [2] YAZDANI D,TOOSI A N, MEYBODJ M R. Fuzzy adaptive artificial fish swarm algorithm[C]//Proc of the 23rd Australasian Joint Conference on Artificial Intelligence. 2010:334-343.
- [3] 王会颖,章义刚.求解聚类问题的改进人工鱼群算法[J].计算机技术与发展,2010,20(3):84-87.
- [4] 姚祥光,周永权,李咏梅,等.人工鱼群与微粒群混合优化算法[J].计算机应用研究,2010,27(6):2084-2086.
- [5] 韦修喜,曾海文,周永权.云人工鱼群算法[J].计算机工程与应用,2010,46(22):26-29.
- [6] WANG Cui-ru,ZHOU Chun-lei,MA Jian-wei. An improved artificial fish-swarm algorithm and its application in feed-forward neural networks[C]//Proc of the 4th International Conference on Machine Learning and Cybernetics. 2005:2890-2894.
- [7] QU Ying,ZHOU Pang. Approximate inference algorithm based on adaptive ant colony and artificial fish swarm algorithm for credal networks[C]//Proc of International Conference on Information Science, Automation and Material System. 2011:656-661.
- [8] SHEN Wei, GUO Xiao-pen, WU Chao, et al. Forecasting stock indices using radial basis function neural networks optimized by artificial fish swarm algorithm[J]. Knowledge-based Systems, 2011, 24(3):378-385.
- [9] 黄华娟,周永权.改进型人工鱼群算法及复杂函数全局优化方法[J].广西师范大学学报,2008,26(1):194-197.
- [10] JIANG Hai-ming,XIE Kang,WANG Ya-fei. Optimization of pump parameters for gain flattened Raman fiber amplifiers based on artificial fish school algorithm[J]. Optics Communications, 2011, 284(23):5480-5483.
- [11] 陈广州,徐晓春,汪家权,等.改进的人工鱼群算法在水质参数识别中的应用[J].水力发电报,2010,29(2):108-113.
- [12] 王晔,吴小俊,王士同,等.基于改进人工鱼群算法的RBF网络及其在人脸表情识别中的应用[J].计算机应用研究,2008,25(9):2643-2646.
- [13] 苏锦旗,吴慧欣,薛惠锋.基于人工鱼群算法的聚类挖掘[J].计算机仿真,2009,26(2):147-150.
- [14] 陈俊清,朱文兴.基于人工鱼群算法的分类规则发现[J].福州大学学报,2007,35(1):25-30.
- [15] 李菲菲,姚坤,刘希玉.一种多微粒群协同进化算法[J].计算机工程与应用,2007,43(22):44-46.
- [16] ZHOU Yong-quan,HUANG Hua-juan,ZHANG Jun-li. Hybrid artificial fish swarm algorithm for solving III-conditioned linear systems of equations[C]//Proc of International Conference on Intelligent Computing and Information Science. 2011:656-661.
- [17] 延丽平,曾建潮.利用多群体PSO算法生成分类规则[J].计算机工程与科学,2007,29(1):94-96.