

基于最小二乘支持向量机的耕地面积预测研究^{*}

周子英, 段建南, 李 萍, 易嗣鑫

(湖南农业大学 资源环境学院, 长沙 410128)

摘 要: 针对耕地面积数据的小样本、复杂非线性特点, 提出一种基于最小二乘支持向量机的耕地面积预测方法。采用相关系数法选择耕地面积的影响因子, 通过粒子群优化算法对最小二乘支持向量机参数进行优化, 最后建立耕地面积与影响因子之间复杂的非线性关系模型。采用湖南省耕地面积数据对模型性能进行验证, 结果表明, 相对于参比模型, 最小二乘支持向量机提高了耕地面积的预测精度, 是一种有效的耕地面积预测方法。

关键词: 最小二乘支持向量机; 粒子群算法; 耕地面积; 影响因子

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2012)03-0873-04

doi:10.3969/j.issn.1001-3695.2012.03.020

Cultivated land prediction based on least square support vector machines

ZHOU Zi-ying, DUAN Jian-nan, LI Ping, YI Si-xin

(College of Resources & Environment, Hunan Agricultural University, Changsha 410128, China)

Abstract: For cultivated land area has small sample data, complex nonlinear characteristic, this paper proposed a cultivated land area method based on least squares support vector machines. Firstly, it selected the influence factors of cultivated land area by correlation coefficient method, then optimized the parameters of least square support vector machines by particle swarm optimization algorithm, lastly built the complex nonlinear model between cultivated area and influence factors. It tested the proposed model by Hunan province cultivated land data, the results show that the proposed model improve the prediction accuracy of cultivated land area compare with other cultivated land area prediction method; the proposed model is an effective prediction method for cultivated land area.

Key words: least square support vector machines(LSSVM); particle swarm optimization(PSO); cultivated land area; influence factors

0 引言

随着我国人口的不断增加、工业化和城镇化的快速推进, 耕地资源短缺越来越严重, 人地矛盾十分尖锐^[1]。科学地对耕地面积进行预测, 制定有效的耕地保护政策, 对我国经济发展和社会稳定有着重要的意义。

多年来, 我国学者对耕地面积预测进行了大量的研究, 将时间序列预测法、回归预测法和灰色预测法等传统模型^[2~6]用于耕地面积预测。由于耕地面积受到国家耕地保护政策、城市发展水平、地区经济、产业结构等多种因素影响, 是一种非线性变化系统, 这些传统的预测模型不能很好地捕捉耕地面积变化过程中的非线性变化特征, 导致预测精度与实际要求存在一定的差距。20 世纪 80 年代以来, 具有自组织、自学习能力的人工神经网络的迅速发展为非线性系统预测开拓了新的研究空间, 其可对任意非线性函数无限逼近, 在耕地预测研究中得到了广泛应用^[7,8]。由于神经网络模型是基于经验风险最小化原则的机器学习方法, 要求数据样本大, 但耕地面积数据属于小样本数据, 往往难以满足大样本这一要求, 因此在实际应用过程中, 易出现过拟合、泛化能力差等现象。最小二乘支持向量机(LSSVM)是一种基于结构风险最小化原则的机器学习方

法, 专门针对小样本、非线性、维数灾难等难题, 其泛化推广能力优异, 较好地克服了神经网络模型过拟合和局部极小等缺陷, 已成功应用于模式识别、回归分析、时间序列预测等领域^[9,10]。

耕地面积变化是一种多变量、非线性的复杂系统, 样本数据有限, 难以用传统数学模型进行描述, 而 LSSVM 不要求对数据分布性质作任何假设, 能够有效地处理多变量、非线性的小样本数据, 适合耕地面积的建模与预测。为了提高耕地面积预测精度, 本文提出了一种基于粒子群算法(PSO)优化 LSSVM 的耕地面积预测模型(PSO-LSSVM), 并通过实例对该模型性能进行验证。

1 LSSVM 耕地面积预测模型

1.1 LSSVM 回归原理

LSSVM 是 Suykens 等人在标准支持向量机的基础上提出的一种机器学习方法, 将标准支持向量机中的损失函数设定成误差平方和, 把不等式约束改成等式约束, 减少待定参数; 同时, 将求解二次规划的问题转换成线性 KKT(Karush Kuhn Tucker)方程组的求解, 简化模型参数确定难度和参数优化复杂度, 可提高模型适应性和精度, 拓宽了支持向量机的应用空间^[11]。

收稿日期: 2011-08-20; **修回日期:** 2011-09-30 **基金项目:** 国家自然科学基金资助项目(41171176); 湖南省教育厅重点资助项目(11A046); 湖南省教育厅资助项目(CX2010B298); 湖南省科技厅资助项目(2009NK3102)

作者简介: 周子英(1974-), 女, 湖南辰溪人, 博士研究生, 主要研究方向为人工智能、土地资源利用与管理(mfkm978@sohu.com); 段建南(1957-), 男, 山西晋城人, 教授, 博导, 主要研究方向为土地资源管理与 3S 集成。

对于训练样本集 $\{(x_i, y_i) \mid (i = 1, 2, \cdots, n), x_i \in R^n, y_i \in R,$
 x_i 和 y_i 分别表示样本输入和输出,通过非线性映射函数 $\varphi(\cdot)$ 将样本映射到高维特征空间,并在高维特征空间进行线性回归,即

$$f(x) = w^T \varphi(x) + b \tag{1}$$

其中: w 为权值向量, b 为偏置量。

根据结构风险最小化原则,式(1)问题求解的 LSSVM 回归模型为

$$\begin{aligned} \min \quad & \|w\|^2 + \frac{1}{2} \gamma \sum_{i=1}^n \xi_i^2 \\ \text{s.t.} \quad & y_i - w^T \varphi(x_i) + b = e_i \end{aligned} \tag{2}$$

其中: γ 为正规化参数, e_i 为实际值与回归函数间的误差。

通过引入拉格朗日乘子 (Lagrange multiplier) 将上述约束优化问题转变为无约束对偶空间优化问题,即

$$\begin{aligned} L(w, b, \xi, \alpha) = & \frac{1}{2} w^T w + \frac{1}{2} \gamma \sum_{i=1}^n \xi_i^2 + \\ & \sum_{i=1}^n \alpha_i (w^T \varphi(x_i) - b + \xi_i - y_i) \end{aligned} \tag{3}$$

其中: α_i 为拉格朗日乘子,按照优化条件

$$\frac{\partial L}{\partial w} = 0, \frac{\partial L}{\partial b} = 0, \frac{\partial L}{\partial \xi_i} = 0, \frac{\partial L}{\partial \alpha_i} = 0$$

可得

$$\begin{aligned} w = & \sum_{i=1}^n \alpha_i \varphi(x_i), \sum_{i=1}^n \alpha_i = 0, \alpha_i = c \xi_i \\ w \varphi(x_i) + b + \xi_i - y_i = & 0 \end{aligned} \tag{4}$$

根据 Mercer 条件,核函数定义为 $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, LSSVM 回归模型为

$$f(x) = \sum_{i=1}^n \alpha_i K(x_i, x_j) + b \tag{5}$$

在一般情况下,基于径向基核函数的 LSSVM 性能要优于其他核函数,因此本文选择径向基核函数作为 LSSVM 核函数。径向基核函数定义为

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{6}$$

其中: σ 为径向基核函数宽度。

最后得到 LSSVM 回归模型:

$$f(x) = \sum_{i=1}^N \alpha_i \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) + b \tag{7}$$

1.2 LSSVM 参数优化

根据 LSSVM 回归原理可知,基于径向基核函数 LSSVM 模型的学习性能与 γ 、 σ 取值密切相关,要得到最优的 LSSVM 模型,首先要获得最优的 γ 和 σ 参数值。PSO 算法源于对鸟群捕食行为的研究,不同于遗传算法的交叉和变异,而是粒子在解空间追踪当前最优粒子进行搜索,其性能通常要优于遗传算法,因此本文采用 PSO 对 LSSVM 参数进行优化。将 LSSVM 参数组 (γ 和 σ) 作为一个粒子,采用耕地面积预测精度作为粒子群的适应度函数,并通过不断迭代搜寻最优值,从而获得 LSSVM 模型最优的 γ 和 σ 。PSO 对 LSSVM 参数优化的具体流程如图 1 所示。

1.3 PSO-LSSVM 的耕地面积预测步骤

PSO-LSSVM 的耕地面积预测模型充分利用了 LSSVM 强大的非线性、高维处理能力,并采用 PSO 优异的全局寻优能

力,建立耕地面积与影响因子之间复杂的非线性关系模型,挖掘出耕地面积的变化规律,从而对耕地面积变化趋势进行预测。耕地面积预测步骤如下:

- a) 收集连续若干年的耕地面积和相应影响因子数据。
- b) 计算影响因子与耕地面积数据间的相关系数,利用相关系数的大小对影响因子进行排序,根据相关系数选择耕地面积影响因子。
- c) 将耕地数据分为训练集与测试集两部分。训练集用于耕地面积预测模型训练和建模,测试集用于检验已建立的耕地面积预测模型性能。
- d) 将耕地面积影响因子作为 LSSVM 输入,耕地面积作为期望输出进行学习, LSSVM 参数通过 PSO 进行优化,从而获得最优 PSO-LSSVM 耕地面积预测模型。
- e) 采用最优 PSO-LSSVM 耕地面积预测模型对测试集进行预测,将预测结果与实际调查耕地面积间的误差进行分析,对模型性能进行检验与评价。

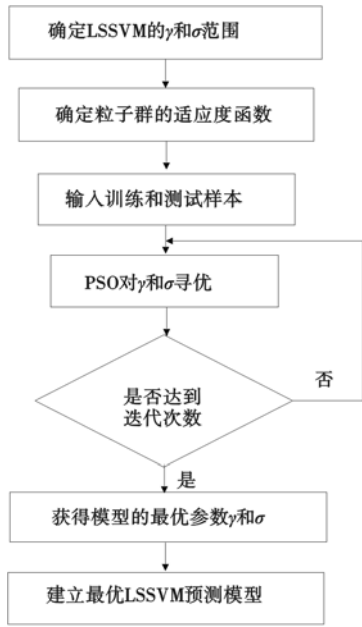


图1 PSO对LSSVM参数的优化流程

1.4 对比模型及评价指标

为了验证 PSO-LSSVM 的耕地面积模型预测性能优劣,选择多元线性回归模型 (MLR)、BP 神经网络模型 (BPNN) 和遗传算法优化的 LSSVM 模型 (GA-LSSVM) 作为对比模型,采用均方误差 (mean squared error, MSE) 和平均绝对误差百分率 (mean absolute percentage error, MAPE) 作为模型性能评价指标。MSE 和 MAPE 定义分别为

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n} \tag{8}$$

$$MAPE = \frac{\sum |y_i - \hat{y}_i| / y_i}{n} \times 100 \tag{9}$$

其中: y_i 为实际值; $\hat{y}_i$ 为预测值; n 为预测样本数。

2 PSO-LSSVM 在耕地面积预测中的应用

2.1 耕地面积数据来源

耕地面积变化的影响因子较多,根据相关研究成

果^[1,3,12,13]初步选择耕地面积(万公顷, y)影响因子为:地区生产总值(亿元, x_1)、总人口数(万人, x_2)、全社会固定资产投资(亿元, x_3)、城市化水平(x_4)、第三产业(亿元, x_5)、耕地保护政策(x_6)、房地产业产值指数(x_7)。数据资料来源于湖南省历年统计年鉴,如表 1 所示。

表 1 1978—2008 年湖南省耕地面积及影响因子

年份	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1978	146.99	5165.91	20.15	11.47	27.34	3	100	344.54
1979	178.01	5223.05	29.86	11.9	30.19	3	101.7	344.04
1980	191.72	5280.95	32.20	12.71	33.59	3	112.3	342.48
1981	209.68	5360.05	46.43	13.4	38.61	3	164.4	342.13
1982	232.52	5452.12	60.66	14.2	42.02	3	181.3	341.74
1983	257.43	5509.43	63.54	14.6	46.27	3	233	339.98
1984	287.29	5561.32	75.7	15.3	54.67	3	256.1	337.14
1985	349.95	5622.49	83.52	16.29	75.15	3	329.1	334.17
1986	397.68	5695.73	99.26	16.91	89.09	3	332	333.05
1987	469.44	5782.61	116.39	17.35	109.90	3	348.6	331.82
1988	584.07	5915.68	140.04	17.65	145.76	3	381.7	332.64
1989	640.80	6013.62	114.41	17.45	168.34	3	331	331.86
1990	744.44	6110.89	124.17	17.55	215.37	3	360.8	331.23
1991	833.30	6166.33	157.07	18.61	250.33	3	382.8	331.02
1992	986.98	6207.78	233.39	19.62	325.90	3	423.4	330.77
1993	1244.71	6245.58	320.24	19.3	390.98	3	525	328.76
1994	1650.02	6302.58	420.89	21.53	527.41	3	588	326.66
1995	2132.13	6392.00	524.01	24.27	676.16	5	674.4	324.97
1996	2540.13	6428.00	678.33	24.99	826.09	5	789.7	323.5
1997	2849.27	6465.00	700.73	25.19	951.73	8	870.2	322.74
1998	3025.53	6502.00	848.59	25.9	1074.14	8	985.9	321.9
1999	3214.54	6532.00	943.34	26.39	1243.30	8	1156.3	321.32
2000	3551.49	6562.05	1066.27	29.75	1473.39	8	1262.9	392.16
2001	3831.90	6595.85	1210.63	30.8	1593.35	8	1380.3	391.26
2002	4151.54	6628.50	1355.87	32	1780.79	8	1534.9	389.1
2003	4659.99	6662.80	1557.00	33.5	1995.78	8	1673	383.37
2004	5641.94	6697.70	1981.29	35.5	2294.60	8	1770	381.65
2005	6511.34	6732.10	2563.96	37	2659.55	8	1865.6	381.6
2006	7568.89	6768.10	3242.39	38.7	3084.96	8	2014.8	378.76
2007	9200.00	6805.70	4294.36	40.45	3657.04	8	2325.1	378.9
2008	11156.64	6845	4940	42.15	4216.16	8	2529.7	378.94

2.2 数据归一化处理

LSSVM 的核函数值通常依赖特征向量的内积,耕地面积影响因子量纲的不同对 LSSVM 模型学习速度产生不利影响。为了提高学习效率,对耕地面积影响因子原始值进行归一化处理,将它们缩放到 $[0,1]$ 之间,具体归一化公式为

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{10}$$

其中: x_i 是原始耕地面积影响因子值; x_i' 为归一化后的因子值; $x_{\max}$ 、 $x_{\min}$ 分别代表每一个影响因子的上界与下界。

2.3 耕地面积影响因子的相关分析

设 Y 为耕地面积集合, X 为一影响因子集合,其数值分别为 $Y = (y_1, y_2, \cdots, y_n)$ 和 $X = (x_1, x_2, \cdots, x_n)$,则该影响因子与

耕地面积间的相关系数计算式为

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \tag{11}$$

其中: r 为相关系数; $\bar{x}$ 、 $\bar{y}$ 为影响因子和耕地面积的平均数。

影响因子的相关系数如表 2 所示。从表 2 可知,初步选择的七个影响因子均与耕地面积成正相关,所有影响因子与耕地面积的相关程度均大于 0.05,结果表明所有耕地面积的影响因子都应该选择。按照相关系数的大小对影响因子进行排序,依次为全社会固定资产投资(x_3)、第三产业(x_5)、地区生产总值(x_1)、房地产业产值指数(x_7)、城市化水平(x_4)、耕地保护政策(x_6)和总人口数(x_2)。

表 2 各影响因子与耕地面积相关系数

内容	x_1	x_2	x_3	x_4	x_5	x_6	x_7
相关系数	0.953	0.661	0.997	0.903	0.970	0.721	0.938
因子排序	3	7	1	5	2	6	4

2.4 结果与分析

2.4.1 各模型的拟合性能分析

1978—2003 年数据输入 PSO-LSSVM 模型学习,对 1978—2003 年数据进行拟合,采用 PSO 对 LSSVM 参数进行寻优,同样采用 MLR、BPNN、GA-LSSVM 分别建立耕地面积预测模型。各模型拟合结果如图 2 所示。

根据图 2 拟合的对比结果可知,MLR 耕地面积的拟合精度平均值为 97.47%,达到预测模型的基本要求,但是由于 MLR 是基于线性建模,因此其拟合精度有待提高。相对于 MLR 模型,BPNN 模型提高了耕地面积预测拟合度,拟合精度平均值达到 98.43%,LSSVM 拟合效果较 MLR 和 BPNN 拟合效果更好,能很好地拟合耕地面积变化与影响因子间的高度非线性关系。同时可知,PSO-LSSVM 模型的拟合结果优于 GA-LSSVM 模型,其平均拟合精度达到了 99.31%。结果表明,PSO 获得的最优 LSSVM 参数要好于 GA,表明 PSO-LSSVM 是一种有效的耕地面积预测方法。

2.4.2 各种模型预测性能对比

评价一个预测模型性能的好坏,主要考察其预测能力而非拟合结果。为进一步检验各模型的预测性能,将数据集分为两部分:1978—2003 年的数据作为训练集,2004—2008 年的数据作为独立测试集,对模型的预测性能进行测试。为避免单个样本预测的偶然性,所有预测模型均采用一步预测法进行预测,在预测第 i 年时,前 $i-1$ 年数据作为训练样本参与建模训练,其后续若干年的数据不得参与建模训练;在预测第 $i+1$ 个样本时,将第 i 年数据加入训练样本集。几种模型对耕地面积独立测试集的预测结果如表 3 和图 3 所示。

表 3 几种模型预测结果精度分析

模型	MSE	MAPE	模型	MSE	MAPE
MLR	122.93	2.80	GA-LSSVM	39.54	1.49
BPNN	73.56	2.10	PSO-LSSVM	9.84	0.79

从表 3 以及图 3 可知,对于独立测试集,PSO-LSSVM 的 MSE 为 9.84,MAPE 为 0.79,MLR 模型预测结果最差,其预测

误差最大,该模型不能很好地表达非线性的耕地面积变化系统;BPNN 的预测精度指标均高于 MLR 模型,但结构设计和参数确定困难,并且十分不稳定,很容易出现过拟合现象;LSSVM 相对于 BPNN,其预测精度有了很大程度的提高,这表明 LSSVM 适合于小样本耕地面积数据的预测,很好地克服了 BPNN 的过拟合、局部极小等缺陷,更加全面地描述了耕地面积复杂的非线性变化规律;同时,PSO-LSSVM 预测精度高于 GA-LSSVM,再一次证明了 PSO 对 LSSVM 参数优化要优于 GA。

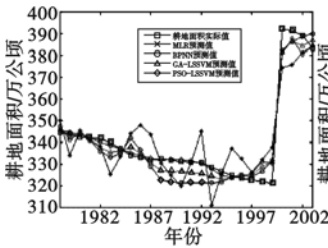


图2 几种模型的拟合结果

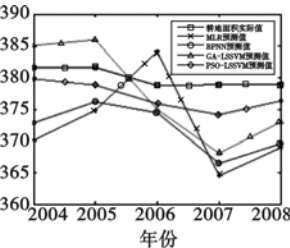


图3 几种模型的预测结果

3 结束语

耕地面积变化受多种因素影响,是一种复杂的非线性系统。传统线性的耕地面积预测模型不能满足实际耕地面积分析与预测要求,神经网络模型虽然具有处理多因素、非线性问题的能力,但其基于经验风险最小原则,对于小样本的耕地面积数据,不能保证模型具有很好的泛化推广性。LSSVM 模型是一种专门针对小样本、多因素的算法,具有不需要对数据分布性质作任何假设且通用性较强等优点。因此,本文将 PSO-LSSVM 引入到耕地面积预测中,结果表明 PSO-LSSVM 非常适合于小样本的耕地面积建模与预测,各项性能评价指标优于 MLR、BPNN 和 GA-LSSVM 模型,是一种快速、高精度的耕地面积预测方法。

(上接第 872 页)

该算法问题在于如果 Pareto 解集是非凸集合,那么就会很难得到 Pareto 解集中的最优解。因此,算法只局限于解决 Pareto 解集为凸集的多目标 VRPTW 决策问题。对于如何解决 Pareto 解集为非凸集合的多目标 VRPTW 决策问题,将是本文今后进一步研究的方向。

参考文献:

[1] 汪定伟,王俊伟,王洪峰,等. 智能优化方法[M]. 北京:高等教育出版社,2007:198-205.
[2] 马良,朱刚,宁爱兵. 蚁群优化算法[M]. 北京:科学出版社,2008:85-115.
[3] 雷德明,严新平. 多目标智能优化算法及其应用[M]. 北京:科学出版社,2009:12-19.
[4] 徐泽水. 不确定多属性决策方法及应用[M]. 北京:清华大学出版社,2004:201-214.
[5] 李世威,王建强,曾俊伟. 一种模糊偏好排序的多目标粒子群算法[J]. 计算机应用研究,2011,28(2):477-480.
[6] 万旭,林健良,杨晓伟. 改进的最大—最小蚂蚁算法在有时间窗车辆路径问题中的应用[J]. 计算机集成制造系统,2005,11(4):572-576.

参考文献:

[1] 王群. 城市化进程中土地资源持续利用问题[J]. 中国土地科学,2003,17(2):47-51.
[2] 杨东,刘强. 基于 GM(1,1)模型的河西地区城镇化与耕地利用变化研究——以张掖市甘州区为例[J]. 西北师范大学学报:自然科学版,2010,46(3):111-115.
[3] 邵晓梅,杨勤业,张洪业. 山东省耕地变化趋势及驱动力研究[J]. 地理研究,2001,20(3):298-306.
[4] 李玉平,蔡运龙. 浙江省耕地变化与粮食安全的分析及预测[J]. 长江流域资源与环境,2007,16(4):466-470.
[5] 赵永华,何兴元,胡远满,等. 岷江上游汶川县耕地变化及驱动力研究[J]. 农业工程学报,2006,22(2):94-97.
[6] 李景刚,何春阳,史培军,等. 近 20 年中国北方 13 省的耕地变化与驱动力[J]. 地理学报,2005,59(2):274-282.
[7] 李立辉,胡月明,陈联诚. 人工神经网络模型在种植业劳均耕地面积预测中的应用[J]. 系统工程理论与实践,2000(9):130-133-140.
[8] 胡喜生,洪伟,吴承祯. 基于 BP 神经网络的福建省耕地预测模型[J]. 福建农林大学学报:自然科学版,2008,37(4):425-427.
[9] 张明光,阎威武,李战明. 基于支持向量机的非线性系统辨识研究[J]. 计算机应用研究,2006,23(5):47-48,82.
[10] 向昌盛,周子英,武丽娜. 基于 ARIMA 和 DSVM 的松毛虫发生面积预测[J]. 湖南农业大学学报:自然科学版,2010,36(4):430-433.
[11] 向昌盛,周子英. 基于支持向量机的害虫多维时间序列预测[J]. 计算机应用研究,2010,27(10):3694-3697.
[12] 张占录,殷秀云. 基于灰色系统及趋势测算的有效耕地面积计算方法[J]. 农业工程学报,2011,27(3):288-294.
[13] 杨柳,刘艳芳. 将微粒群和支持向量机用于耕地驱动因子选择的研究[J]. 武汉大学学报:信息科学版,2010,35(2):248-251.
[7] 刘志硕,中金升,柴跃廷. 基于自适应蚁群算法的车辆路径问题研究[J]. 控制与决策,2005,20(5):562-566.
[8] MOHAMMAD A, MIGUEL M. Application of an ant algorithm for layout optimization of tree network[J]. Engineering Optimization, 2006,38(3):353-369.
[9] CHEN C H, TING C J. An improved ant colony system algorithm for the vehicle routing problem[J]. Journal of the Chinese Institute of Industrial Engineers,2006,23(2):115-126.
[10] BAHREININEJAD A, HESAMFAR P. Subdomain generation using emergent ant colony optimization[J]. Computers and Structures, 2006,4(5):1719-1728.
[11] ISSMAIL E, CALAMAI P, BASIR O. Exchange strategies for multiple ant colony system[J]. Journal of Information Sciences,2006,3(1):46-63.
[12] SIQUEIRA P H, STEINER M T, SCHEER S. A new approach to solve the traveling salesman problem[J]. Neurocomputing,2006,70(4):1013-1021.
[13] BIANCHI L, KNOWLES J, BOWLER J. Local search for the probabilistic traveling salesman problem: correction to the 2-p-opt and 1-shift algorithms[J]. European Journal of Operational Research, 2005,162(1):206-219.