# 基于改进遗传算法的支持向量机预测模型研究

陈锦青，韩延杰
（福州大学 管理学院 信息管理与信息系统专业，福建 福州 350108）

**摘 要**：作为一种新的机器学习方法，支持向量机的参数选择没有一个统一的模式和标准。为了克服这一缺点，对遗传算法进行改进，构造一种混沌云自适应模拟退火遗传算法（CCASAGA）对支持向量机回归参数进行优化。该算法将混沌优化、基于云模型的自适应控制机制和模拟退火的Metropolis准则结合起来，并采取精英保持策略加快算法的收敛速度。利用改进后的CCASAGA-SVR预测模型对某股份制银行ATM机现金需求进行预测，并引入GA-SVR模型和BP神经网络模型进行对比，从而证实该预测模型具有更高的预测精度。

**关键词**：遗传算法；支持向量机；BP神经网络；预测
**中图分类号**：TP391 **文献标识码**：A **文章编号**：1674-7720(2013)24-0082-03

## Support vector machine prediction model based on improved genetic algorithm

Chen Jinqing, Han Yanjie
(Department of Information Management and System Program, Fuzhou University, Fuzhou 350108, China)

**Abstract**: As a new method of machine learning-support vector machine(SVM), there is not an unified mode and standards to select parameters. In order to overcome this shortcoming, the author improves the genetic algorithm, and proposes a chaos Cloud-based adaptive simulated annealing genetic algorithm (CCASAGA) to optimize the parameters of support vector regression machine. This method combines chaos optimization, adaptive control mechanism based on Cloud model and simulated annealing Metropolis criterion, and also takes the elite hold strategy to accelerate the speed of convergence of the algorithm. Finally, take the constructed CCASAGA-SVR model to predict the ATM cash demand of a joint-stock bank. To illustrate the proposed method has a higher prediction accuracy, this paper introduces standard GA-SVR model and BP neural network model as a comparison.

**Key words**: genetic algorithm; support vector machine; BP neural network; forecasting

支持向量机SVM(Support Vector Machine)最初于20世纪90年代由Vapnik等人首先提出，是一种基于统计学习理论的新型的通用学习方法[1]。支持向量机模型可以用于分类和预测。在应用支持向量机模型进行预测时，对预测精度产生重要影响的参数是：惩罚因子$C$、不敏感损失函数参数$\varepsilon$和核函数及其参数。因此，本文的预测模型需要利用遗传算法对这3个参数进行优化以提高SVR预测模型的预测精度。标准遗传算法SGA(Standard Genetic Algorithm)存在早熟收敛和易陷入局部最优解的缺陷，主要表现在种群中所有个体状态趋于一致而停止进化，算法不能找到令人满意的解。针对上述缺点，各国学者对遗传算法的编码方式、适应度函数的设计、遗传算子机理等进行深入研究，提出了众多的改进方法，包括免疫遗传算法[2]、基于多样化成长策略的遗传算法[3]和模糊自适应遗传算法[4]等。

本文在前人研究成果的基础上，提出一种混沌云自适应模拟退火遗传算法CCASAGA(Chaos Cloud-based Adaptive Simulated Annealing Genetic Algorithm)，使用混沌映射优化遗传算法的初始种群，采用云模型实现交叉概率和变异概率的自适应调整，引入模拟退火避免算法陷入局部最优，并采取精英保持策略，防止进化过程中产生的优秀个体模式遭到破坏，提高了算法的收敛速度。运用CCASAGA对SVR的参数进行优化，寻找到更好的参数值，从而提高模型的预测精度。

## 1 遗传算法的改进

### 1.1 利用混沌映射优化遗传算法初始种群

传统遗传算法的初始种群由随机的方法产生，因此每次寻优效果可能不尽相同，且容易导致算法陷入局部

最优。本文采用混沌映射优化遗传算法的初始种群，利用混沌变量具有的遍历性、随机性和内在规律性，在一定范围内不重复地遍历所有状态，从而保证种群分布均匀，具有多样性。

本文采用 Logistic 映射产生混沌变量。Logistic 映射公式如下[5]：

$$x_{n+1}=4\times x_n\times(1-x_n)\quad n=0,1,2,\cdots \tag{1}$$

式中，$x_n$ 为经第 $n-1$ 次混沌迭代后产生的混沌变量，当 $x_n\in(0,1)$ 且 $x_n\notin\{0.25,0.5,0.75\}$ 时，系统将完全处于混沌状态。

对变量 $x_i\in(a_i,b_i)$，$i=1,2,3,\cdots,n$ 进行 Logistic 混沌映射的过程如下：

(1)将变量 $x_i$ 从定义空间 $S_0$ 线性映射到空间 $S_1$：$\{(0,1),i=1,2,3,\cdots,n\}$，变换公式为：

$$x_i'=(x_i-a_i)/(b_i-a_i),i=1,2,3,\cdots,n \tag{2}$$

(2)令 $x_n=x_i'$，由式(1)进行混沌迭代产生下一代混沌变量 $x_i^{(1)}$；

(3)将混沌变量 $x_i^{(1)}$ 从空间 $S_1$ 线性映射回定义空间 $S_0$，得到新参数：

$$x_i^{(2)}=a_i+x_i^{(1)}(b_i-a_i),i=1,2,3,\cdots,n \tag{3}$$

本文利用 Logistic 映射对初始种群的优化方法如下：

(1)将遗传算法的初始种群 $POP^{(1)}$ 中每个个体通过混沌映射为其定义空间中的新个体，由此得到新种群 $POP^{(2)}$；

(2)比较映射前后 $POP^{(1)}$ 和 $POP^{(2)}$ 中个体适应度值。如果 $POP^{(2)}$ 中个体适应度值大于 $POP^{(1)}$ 中相应的个体适应度值，则用 $POP^{(2)}$ 新个体代替 $POP^{(1)}$ 对应个体；否则不进行替代；

(3)计算 $POP^{(1)}$ 中最优个体适应度值。如果连续 $M$ 次搜索后种群最优适应度值保持不变，则以当前种群为找到的初始最优种群；否则，转至步骤(1)继续搜索。

### 1.2 基于云模型的遗传算子自适应调整

云模型是我国李德毅院士提出的一种用自然语言值表示的定性概念与其定量数据表示之间的不确定性转换模型，同时具有模糊性和随机性，为定性与定量相结合的信息处理提供了有力手段[6]。由此，将云模型引入遗传参数自适应调整过程，可以更好地实现其用自然语言描述的作用机理。本文利用 X 条件云发生器生成交叉概率和变异概率，使它们既有传统自适应遗传算法的趋势性，加快算法收敛速度，又具有随机性，使最大适应度的个体交叉和变异概率不为零，从而提高算法跳出局部最优的能力。具体算法如下：

(1)自适应交叉概率 $P_c$ 新算法

$$E_x=f_{avg},\qquad E_n=(f_{max}-f_{avg})/c_1$$

$$H_e=E_n/c_2,\qquad E_n'=\mathrm{randn}(E_n,H_e)$$

$$P_c=\begin{cases}k_1 e^{\frac{-(f'-E_x)^2}{2(E_n')^2}}, & f'\geqslant f_{avg}\\ k_3, & f'<f_{avg}\end{cases} \tag{4}$$

(2)自适应变异概率 $P_m$ 新算法

$$E_x=f_{avg},\qquad E_n=(f_{max}-f_{avg})/c_3$$

$$H_e=E_n/c_4,\qquad E_n'=\mathrm{randn}(E_n,H_e)$$

$$P_m=\begin{cases}k_2 e^{\frac{-(f-E_x)^2}{2(E_n')^2}}, & f\geqslant f_{avg}\\ k_4, & f<f_{avg}\end{cases} \tag{5}$$

交叉概率 $P_c$ 和变异概率 $P_m$ 计算公式中参数的含义与 AGA 中的一样，$\mathrm{randn}(E_n,H_e)$ 表示生成期望为 $E_n$、标准差为 $H_e$ 的正态随机数。$E_n$ 影响云的陡峭程度，$E_n$ 越大，则云覆盖的水平宽度越大，从而使更多较优个体取到较小的交叉变异概率，根据“$3E_n$”规则，$c_1$ 和 $c_3$ 取 3.0 附近的值。$H_e$ 则决定了云滴的离散程度，$H_e$ 过小，会在一定程度上丧失随机性，$H_e$ 过大，又会在一定程度上丧失稳定倾向性。建议 $c_2$ 和 $c_4$ 取[5，15]之间的值。

### 1.3 引入模拟退火思想

在遗传算法中引入模拟退火思想，具体来说就是根据 SA 中的 Metropolis 准则来判断是否接受由交叉和变异操作产生的新个体代替交叉和变异前的旧个体。在 Metropolis 准则指导下，算法在接受优质解的同时，也有限度地接受较差解，增强了算法局部搜索能力，确保种群能始终朝着最优种群的方向进化。

对于目标函数求最小的优化问题，SA 接受新解的概率为：

$$P_{acc}=\begin{cases}1, & f(x')<f(x)\\ \exp(-(f(x')-f(x))/T), & f(x')\geqslant f(x)\end{cases} \tag{6}$$

其中，$x$ 为当前解，$x'$ 为新解，$f(\cdot)$ 表示解的目标函数值，$T$ 为温度。

## 2 基于 CCASAGA 的 SVR 参数优化设计

对遗传算法进行改进的目的是为了更好地对支持向量机的参数进行优化，以取得更好的预测效果。利用 CCASAGA 优化 SVR 参数的流程如图 1 所示。

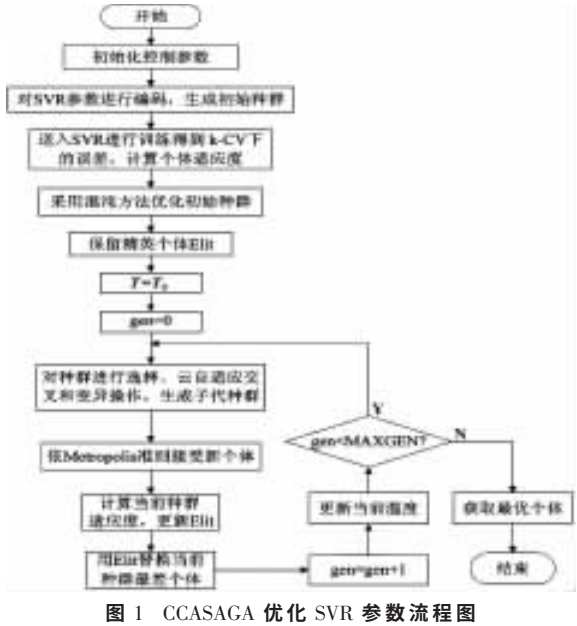


图 1 CCASAGA 优化 SVR 参数流程图

## 3 实验结果

### 3.1 实验数据

本文数据来源于某股份制银行某市分行 ATM 交易后台数据库中的取款数据。这里选取某台 ATM 机 2010 年 12 月 6 日~2012 年 9 月 13 日的日取款量数据作为研究样本，样本量的大小为 648，以前 643 天数据为训练集，后 5 天数据为测试集。模型训练完后对后 5 天的 ATM 现金需求量进行多步预测，并在测试集上检验模型的预测效果。

### 3.2 实验的实现和结果

整个模型算法通过 MATLAB 软件编程实现，其中支持向量机的实现用到了台湾大学林智仁教授团队所编写的 Libsvm-3.1 工具箱，对遗传算法的改进用到了英国谢菲尔德大学开发的 GATBX 遗传算法工具箱。整个模型的参数设置如表 1 所示。

表 1 模型参数取值表

| 项目 | 参数名称 | 参数取值或搜索范围 |
|---|---|---|
| SVR 涉及参数及搜索范围 | 惩罚因子 $C$ | (0，100] |
| | 径向基核函数参数 $\gamma$ | [0，1000] |
| | 损失函数参数 $\varepsilon$ | [0.01，1] |
| | 交叉验证折数 $k$ | 5 |
| CCASAGA 涉及参数 | 种群规模 SizePop | 30 |
| | 最大进化代数 MAXGEN | 100 |
| | 混沌优化连续迭代次数 $M$ | 20 |
| | 模拟退火初始温度 $T_0$ | 1 000 |
| | 云模型控制参数 | $k_1=k_2=1$，$k_3=k_4=0.5$，$c_1=c_3=3$，$c_2=c_4=10$ |

应用前面提出的预测模型对 ATM 机的现金需求进行预测。首先运用改进后的遗传算法对 SVR 参数进行优化，寻优过程如图 2 所示。最终得到最佳的 $C=0.270\,589$，$\gamma=9.264\,51$，$\varepsilon=0.081\,102\,7$。将最佳参数代入 SVR 模型进行训练，得到最优的 SVR 预测模型，用最优模型预测后 5 天的现金需求量，并与测试集数据进行对比，得到最终的预测结果，如图 3 所示。

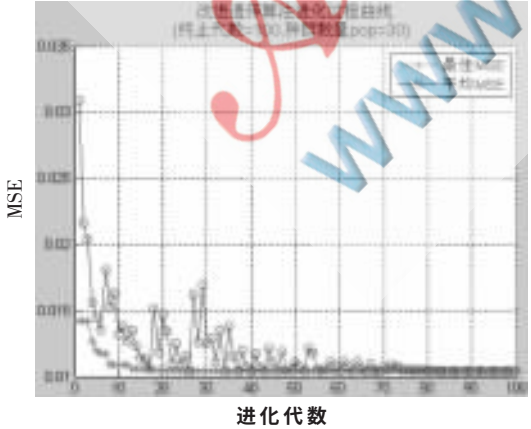


图 2 改进遗传算法对 SVR 参数寻优过程曲线

将本文提出的 CCASAGA-SVR 模型与传统 GA-SVR 模型及 BP 神经网络模型[5]预测结果进行对比，如表 2 所示。

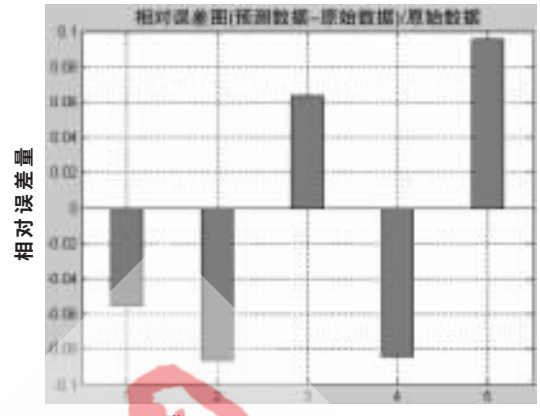


图 3 CCASAGA-SVR 预测相对误差图

由表 2 可知，本文提出的 CCASAGA-SVR 模型预测精度最高，而神经网络模型相对来说预测精度最低。本文对遗传算法改进后，使用 CCASAGA-SVR 的预测误差比 GA-SVR 的预测误差降低了约 1.03 个百分点。然而，本身 GA-SVR 模型的预测平均精度也在 90%以上，这一方面说明采用标准遗传算法对支持向量回归机进行优化也可以得到预测精度较高的预测模型，另外也说明本文对遗传算法的改进是有效的，能够在原来较好的预测效果上进一步提高预测精度。另外，BP 神经网络模型的总体预测精度较差，而且预测误差波动范围较大，这可能是由于其在训练过程中产生了过拟合的现象，因此其泛化能力不如支持向量回归机。

表 2 不同模型预测误差比较

| 预测模型 | MAE | MAPE/% | 最小 APE/% | 最大 APE/% |
|---|---|---|---|---|
| CCASAGA-SVR | 11 518.81 | 7.69 | 5.53 | 8.61 |
| GA-SVR | 13 097.06 | 8.72 | 6.17 | 10.49 |
| BP 神经网络 | 18 062.53 | 12.03 | 7.09 | 15.91 |

## 参考文献

[1] 曹建芳，王鸿斌.一种新的基于 SVM-KNN 的 Web 文本分类算法[J].计算机与数字工程，2010，38(4)：59-61.

[2] 王洁，高家全，方志民，等.一种新的免疫遗传算法及应用[J].计算机应用与软件，2010，27(12)：89-91.

[3] 袁煜明，范文慧，杨雨田，等.一种基于多样化成长策略的遗传算法[J].控制与决策，2009，24(12)：1801-1804.

[4] Guo Yiqiang，Wu Yanbin，Ju Zhengshan，et al.Remote sensing image classification by the chaos genetic algorithm in monitoring land use changes[J].Mathematical and Computer Modelling，2010，51(11)：1408-1416.

[5] 李仿华，王爱平，姚丽娜，等.基于遗传优化的 RBF-BP 网络的实时故障检测[J].微型机与应用，2012，31(8)：

90-93.
[6] 戴朝华,朱云芳,陈维荣.云自适应遗传算法[J].控制理论与应用,2007,24(4):646-650.

(收稿日期:2013-08-27)

作者简介:
陈锦青,女,1989 年生,硕士研究生,主要研究方向:商务智能与数据挖掘。
韩延杰,男,1984 年生,硕士研究生,主要研究方向:商务智能与数据挖掘。