

# 基于智能集成粒子群算法的时间序列数据挖掘研究

张 健

(三江学院计算机科学与工程学院, 南京 210012)

**摘 要:**针对单一算法在处理复杂时间序列数据时存在缺陷以致无法挖掘全部信息的问题,提出了智能集成架构,给出了四种集成结构,并分析了它们的适用情况。针对一类随机噪声干扰的时间序列数据,采用并联嵌套建模结构,提出嵌套双种群粒子群算法的自回归滑动平均(ARMA)模型,用于挖掘数据中的随机性趋势;提出基于概率密度控制(PDF)的最小二乘支持向量机(LSSVM),用于挖掘数据中的确定性趋势,两种模型并联补集成实现对数据信息的充分挖掘。通过一组实验验证了所提方法的效果。

**关键词:**时间序列;支持向量机;智能集成;自回归滑动平均

**中图分类号:**TP399

**文献标志码:**A

## 引 言

随着信息时代的到来,大数据分析已成为各个领域愈来愈重视与依赖的技术手段。其中,时间序列数据广泛存在于农业、金融、制造业等各个领域。时间序列挖掘是从大量的时间序列数据中提取数据中隐含的规律与知识,通过对时间序列数据进行挖掘分析,能够掌握事物的发展规律,从而对其未来趋势进行预测。

时间序列建模与预测方法一般分为传统方法与智能方法两类。传统方法包括线性回归分析<sup>[1-3]</sup>、非线性回归分析<sup>[4-6]</sup>、自回归滑动平均(ARMA)建模<sup>[7-9]</sup>、偏最小二乘法<sup>[10]</sup>、灰色预测<sup>[11-13]</sup>等。

智能方法采用专家系统<sup>[14]</sup>、模糊规则<sup>[15]</sup>、神经网络<sup>[16-17]</sup>、支持向量机<sup>[18-20]</sup>等智能技术实现预测建模。专家系统建模借鉴专家经验知识实现对生产过程的描述,具有非常好的解释性,然而其知识获取存在瓶颈,学习能力差。模糊逻辑与专家系统类似,也是根据专家经验知识实现对生产过程的描述,区别在于它采用模糊推理方法能够很好地处理不确定信息。基于模糊规则的建模技术同样受限于所获取的知识,并且具有模型精度不高的问题。

人工神经网络与支持向量机是两种具有代表性的基于数据的机器学习技术。当样本量足够大时,人工神经网络能够以任意精度逼近工业对象的非线性特性,因此被广泛应用于预测建模问题中。然而,人工神经网络的建模精度依赖于学习样本并且模型训练容易陷入局部最优。支持向量机技术建立在统计学习理论基础之上,它能够依靠有限的样本信息,基于结构风险最小化理论,在模型复杂性和模型学习能力之间寻求最佳折衷,因此具有优良的泛化能力。需要指出的是,人工神经网络与支持向量机技术虽然具有出色的非线性逼近能力,但是它们与传统建模方法一样,建立的都是黑箱模型,其模型精度依赖于所获取的样本信息。

对于复杂的预测问题,采用一种建模方法往往无法达到预测精度,因此需要集成多种建模技术,吸收各方建模优势,从而达到精确预测目的。智能集成建模是将两种或两种以上的建模方法,按一定的方式集成后实现对复杂工业过程建模,其中至少一种为智能建模方法。比如,文献[21]集成粗糙集理论与支持向量机从而建立粗支持向量机模型,实现对时间序列预测;文献[20]针对时间序列数据非线性、冗余特征,集成支持向量机技术与粒子群优化算法实现预测;文献[22]针对数据存在

收稿日期:2015-06-04

作者简介:张 健(1975-),男,江苏淮安人,实验师,硕士,主要从事计算机应用方面的研究,(E-mail)zhang1jian086@126.com

高度的非线性、耦合性和多因素的影响,采用集成遗传算法与最小二乘支持向量机的预测建模方法,从而提高了预测精度;文献[23]提出了一种神经网络和灰色预测相结合的税收预测新方法,与单一方法相比,该方法具有更高的精度。

本文提出了智能集成架构,给出了四种智能集成结构并分析了它们的适用情况。针对一类随机噪声干扰的时间序列数据,采用并联嵌套建模结构,提出嵌套双种群粒子群算法的自回归滑动平均模型,用于挖掘数据中的随机性趋势;提出基于概率密度控制的最小二乘支持向量机,用于挖掘数据中的确定性趋势,两种模型并联补集成实现对数据信息的充分挖掘。通过一组实验验证了所提方法的效果。

### 1 智能集成架构

智能集成是将两种或两种以上的模式挖掘方法,按一定的方式集成后实现对复杂数据规律或模式挖掘,其中至少一种为智能建模方法。智能集成模式挖掘方法的形式与结构主要有四种。

(1) 并联补集成结构。并联补集成结构包括两个子模型,两个模型没有主次之分,且相互之间互为补充。该结构中的两个子模型通常由两种建模方法得到,单一建模方法能够挖掘时间序列数据中的部分信息以获知对应规律,但由于方法所限,无法获知数据中的全部信息,因此依靠两种建模方法互为补充以充分挖掘数据中隐含的规律或模式。

叠加形式分为相加与相乘两种。并联叠加集成结构如图1与图2所示。图中,  $X_1$  为模型1的输入,  $Y_1$  为模型1的输出,  $Y_1 = f_1(X_1)$ 。  $X_2$  为模型2的输入,  $\delta$  为模型2的输出,  $\delta = f_2(X_2)$ 。图1中,  $Y = Y_0 + \delta$ ; 图2中,  $Y = \delta Y_0$ 。

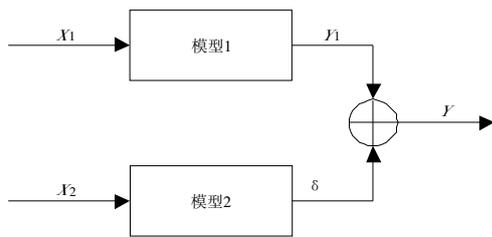


图1 相加形式的并联补结构

(2) 加权叠加集成结构。加权叠加集成结构由多个子模型加权后叠加构成,其中每个子模型对应的权重大小决定了它在集成模型中所起的作用。该结构中的多个子模型通常由多种建模方法得到,单一建模方法能

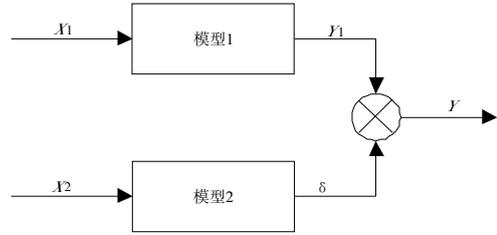


图2 相乘形式的并联补结构

够挖掘时间序列数据中的部分信息以获知对应规律,但由于方法所限,无法获知数据中的全部信息,因此依靠多种建模方法互为补充以充分挖掘数据中隐含的规律或模式。

加权叠加集成结构如图3所示。其中,  $Y = \sum_{i=1}^n \omega_i Y_i, Y_i = f_i(X_i), i = 1, 2, \dots, n$ 。

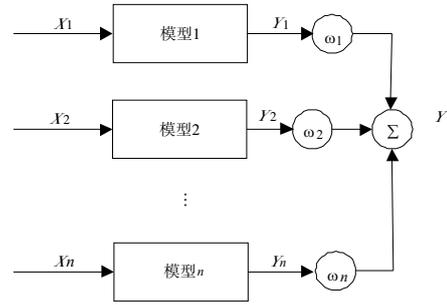


图3 加权并集成结构

(3) 串联集成结构。串联集成结构包括两个甚至更多个子模型,除了第一个和最后一个模型,每一个模型都是前面一个模型的输出,同时也是后一个模型的输入。非线性动态系统通常采用这种形式,比如,采用神经网络反映系统静态时的非线性特性,采用 NARMX(具有外生变量的非线性自回归滑动平均)表征动态特性。串联集成结构如图4所示。

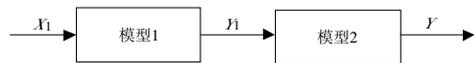


图4 串联集成结构

(4) 模型嵌套集成结构。嵌套集成结构包括至少两个子模型,其中一个称为基模型,用来对工业过程的主体结构进行建模,其它子模型则嵌套在基模型中,用来对基模型中的未知参数建模,如图5所示。比如将蚁群算法<sup>[1-2]</sup>、粒子群优化算法<sup>[3-5]</sup>、遗传算法<sup>[6-7]</sup>等仿生算法应用到系统辨识中,用来实现模型中的参数估计。

### 2 嵌套双种群粒子群算法的 ARMA 模型

ARMA 时间序列模型理论非常完善,对于一个平

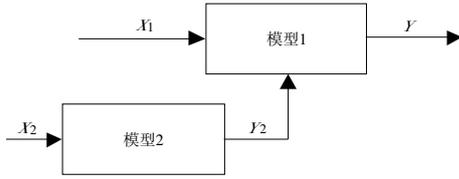


图 5 模型嵌套集成

稳、零均值的时间序列,如果采取合适的阶次与系数,它能保证拟合出的模型预报残差为零均值噪声。

本文提出双种群粒子群优化算法(cPSO),其中一个子群执行自适应网格粒子搜索,以保持种群的多样性,提高算法的全局搜索能力;另外一个子群按照快速收缩粒子群算法搜索,具有非常出色的收敛性能。采用cPSO算法优化确定 ARMA 模型的阶次与系数以最小化模型预报残差。

算法步骤为:

第一步:采用单位根检验法(ADF)对时间序列数据进行平稳性检验,如果序列为零均值平稳序列则直接用于 ARMA 模型建模,否则需要对时间序列数据进行平稳化处理。

第二步:设置模型阶次与系数优化准则为模型预报的均方根误差最小化。

第三步:采用cPSO算法拟合 ARMA 模型系数与阶次。cPSO 算法原理如图 6 所示,其过程为:

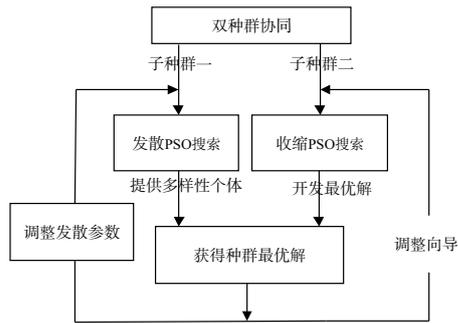


图 6 cPSO 算法流程

(1) 划分网格

将每一维决策变量平均分成  $g_p$  段。

(2) 划分子种群

将整个种群分为两个子群,即发散 PSO 搜索子群(简称网格子群)与收缩 PSO 子群(简称 PSO 子群),两个子种群的粒子个数分别为  $Pop_z$  与  $Pop_s$ 。

(3) 种群初始化

在每一个格子内都随机生成一个粒子  $px_q \in R^2$ ,从而构成发散搜索子群,  $\{px_q | q = 1, \dots, Pop_z\}$ 。另外,在整个决策空间内随机生成  $Pop_s$  个粒子,从而得到 PSO 子

群  $\{px_q | q = Pop_z + 1, \dots, Pop_z + Pop_s\}$ 。令 PSO 子群的初始速度为 0,对于第  $k$  个粒子,  $pbest_q = px_q$ , 即其初始个体向导  $pbest_q$  为其本身,  $q = Pop_z + 1, \dots, Pop_z + Pop_s$ 。对于全局向导  $gbest$ , 利用某个随机选择的粒子的位置对其初始化<sup>[24]</sup>。

(4) 向导调整

根据优化准则即模型预报误差均方根最小化评估每个粒子,得到  $f_{eval}(px_q), q = 1, \dots, Pop_z + Pop_s$ , 并按照式(1)~式(2)调整 PSO 子群的个体向导与全局向导

$$pbest_q \leftarrow \begin{cases} pbest_q & \text{if } f_{eval}(px_q) \geq f_{eval}(pbest_q) \\ px_q & \text{if } f_{eval}(px_q) < f_{eval}(pbest_q) \end{cases} \quad (1)$$

$$gbest \leftarrow \begin{cases} gbest & \text{if } f_{eval}(gbest) \leq \min_{q=1}^{Pop_z+Pop_s} \{f_{eval}(px_q)\} \\ px_i & \text{if } f_{eval}(px_i) = \min_{q=1}^{Pop_z+Pop_s} \{f_{eval}(px_q)\} \wedge f_{eval}(px_i) < f_{eval}(gbest) \end{cases} \quad (2)$$

(5) 发散参数调整

根据  $gbest$  所在格子(这里称为向导网格)的位置调整每个网格粒子的搜索范围。调整后使所有发散 PSO 粒子的搜索范围都包括向导网格区间,如图 7 所示。为了表述简单,假设将每一维决策变量范围分成 3 段,二维空间,因此总共分割成 9 个网格,其中  $gbest$  位于第 5 个网格内。在调整前粒子 1 的搜索范围为网格 1,调整后粒子 1 的搜索范围变为网格 1、2、4、5 构成的空间。同样,在调整前粒子 2 的搜索范围为网格 2,调整后粒子 2 的搜索范围变为网格 2、5 构成的空间。

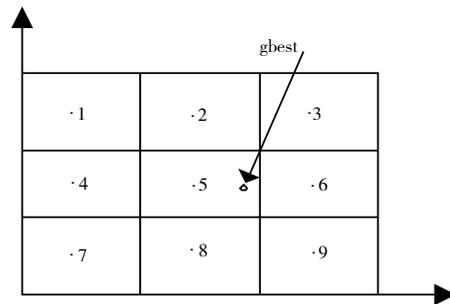


图 7 网格调整实例

(6) 发散 PSO 子群位置更新

发散 PSO 子群的位置  $px_q (q = 1, \dots, Pop_z)$  按照式(3)进行更新

$$px_q \leftarrow rand(\Omega_q^2) \quad (3)$$

其中,  $\Omega_q^2$  为第  $q$  个粒子的网格搜索区间,  $rand(\Omega_q^2)$  表示在  $\Omega_q^2$  区间内随机生成一个新点。

(7) PSO 子群位置更新

收缩 PSO 子群的位置  $px_q (q = Pop_z + 1, \dots, Pop_z +$

Pop<sub>s</sub>) 按照式(4)进行更新

$$\begin{cases} v_q \leftarrow \chi[v_q + c_1 c_3 (pbest_q - px_q) + c_2 c_4 (gbest - px_q)] \\ px_q \leftarrow v_q + px_q \end{cases} \quad (4)$$

其中,  $c_1, c_2$  为加速因子参数,  $\chi = 2 / |2 - (c_1 + c_2) - \sqrt{(c_1 + c_2)^2 - 4(c_1 + c_2)}|$ , 其中  $c_1 + c_2 > 4$ 。  $c_3$  与  $c_4$  为(01)区间内的随机数。

第四步:验证所建立的时间序列 ARMA 模型。

### 3 基于 PDF 的 LSSVM

本文在前人研究的基础上,提出一种新 LSSVM 模型参数选择准则。通过该准则,可以使得 LSSVM 模型的残差 PDF 向给定的高斯分布逼近<sup>[24]</sup>,从而实现数据分析与预测的泛化性与精度提高的目的。其算法原理<sup>[24]</sup>为:

假设回归问题的一组样本数据集为  $D = \{(x_1, y_1), \dots, (x_j, y_j), \dots, (x_l, y_l)\}$ ,  $x_j \in R^n$ ,  $y_j \in R$ , 并且存在一个非线性函数:

$$f(x) = [\omega, \varphi(x)] + b \quad (5)$$

其中,  $\omega \in R^m$  表示权向量,  $b$  表示偏置项,  $[\cdot, \cdot]$  表示点乘,  $\varphi(x): R^n \rightarrow R^m$  表示输入空间向高维特征空间(维数不受限制)的非线性映射。

通过 LSSVM,优化问题可以转化或描述为:

$$\min_{\omega, \zeta} J(\omega, \zeta) = \frac{1}{2} \omega^2 + \frac{1}{2} C \sum_{j=1}^l \zeta_j^2$$

$$s. t. y_j = [\omega, \varphi(x_j)] + b + \zeta_j, j = 1, \dots, l \quad (6)$$

其中,  $\zeta_j \in R$  和  $C$  分别用来表示残差和惩罚系数。

对 Lagrangian 函数进行建立,并依据 KKT (Karush-Kuhn-Tucker) 条件,得到

$$\begin{aligned} \frac{\partial L}{\partial \omega} &= \|\omega\| - \sum_{j=1}^l \alpha_j \varphi(x_j) = 0 \\ \frac{\partial L}{\partial \zeta} &= C \sum_{j=1}^l \zeta_j - \sum_{j=1}^l \alpha_j = C \zeta_j - \alpha_j = 0 \\ \frac{\partial L}{\partial b} &= \sum_{j=1}^l \alpha_j = 0 \\ \frac{\partial L}{\partial \alpha} &= \omega \varphi(x_j) + b + \zeta_j - y_j = 0 \end{aligned} \quad (7)$$

消除  $\omega, \zeta$  后,得到线性方程

$$\begin{bmatrix} 0 & \tilde{I}^T \\ \tilde{I} & \Omega + I/C \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \quad (8)$$

其中,  $\alpha_j \in R$  为 Lagrangian 算子,  $y = [y_1, \dots, y_l]$ ,  $\tilde{I} = [1, \dots, 1]$ ,  $\alpha = [\alpha_1, \dots, \alpha_l]$ , 应用 Mercer 条件:

$$\begin{aligned} \Omega_{jk} &= [\varphi(x_j)^T, \varphi(x_k)] = K(x_j, x_k) \\ j, k &= 1, 2, \dots, l \end{aligned} \quad (9)$$

核函数取高斯径向基函数

$$K(x_j, x_k) = \exp\left(-\frac{\|x_j - x_k\|^2}{2\sigma^2}\right) \quad (10)$$

则,待求 LSSVM 回归模型为:

$$f(x) = \sum_{j=1}^l \alpha_j K(x, x_j) + b \quad (11)$$

其中,求解式(8)可获得  $\alpha_j$  与  $b$  的值。

在前面的模型构建过程中,  $C$  和  $\sigma$  是可调的,它们的值一旦确定,就得到了最小二乘支持向量机的具体模型。

通过文献[20]中用到的标准网格搜索算法求取 LSSVM 模型的参数。

残差  $\xi$  为

$$\xi = y - \left(\sum_{i=1}^l \alpha_i K(x, x_i) + b\right) \quad (12)$$

进一步可以写成如下函数形成

$$\xi = \Pi(x, y, C, \sigma) \quad (13)$$

以  $\gamma_\xi$  表示  $\xi$  的概率密度分布,  $\gamma_\xi$  为  $C$  和  $\sigma$  的函数,即  $\gamma_\xi(x, y, C, \sigma)$ 。可以通过调整  $C$  和  $\sigma$  的数值使  $\gamma_\xi$  接近目标高斯分布。

以  $\gamma_{target}$  表示目标高斯分布的概率密度函数:

$$\gamma_{target}(z) = \frac{1}{\sqrt{2\pi}\sigma_C} e^{-\frac{(z-\mu)^2}{2\sigma_C^2}} \quad (14)$$

定义参数选择准则

$$J = \min_{C, \sigma} H = \min_{C, \sigma} \int_{-\infty}^{+\infty} (\gamma_{target}(z) - \gamma_\xi(z))^2 dz \quad (15)$$

### 4 实验研究

针对随机噪声干扰的时间序列数据,本文采用并联嵌套建模结构,利用子模型嵌套 cPSO 的 ARMA 模型挖掘数据中的随机性趋势;利用基于 PDF 的 LSSVM 挖掘数据中的确定性趋势,两种模型并联补集成实现对数据信息的充分挖掘。

实验以某城市降雨量时间序列数据为收集对象,该数据集受到随机因素影响。分别采用本文所提方法(ARMA - LSSVM)与单纯的 PDF - LSSVM 分别对上述数据分析与预测,给定的目标高斯概率密度函数的均值  $\mu = 0$ , 方差  $\sigma_C^2 = 7.5$ 。图 8 ~ 图 11 分别表示 PDF - LSSVM 与 ARMA - LSSVM 模型在训练结果与样本数据拟合程度、预报结果与测试数据拟合程度、训练残差、预报残差的对比。图 12 ~ 图 13 分别表示对两种模型的残差的自相关以及预报残差的自相关的分析。

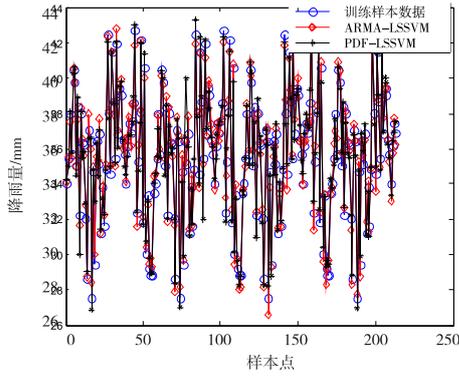


图 8 PDF-LSSVM 与 ARMA-LSSVM 模型对训练结果与样本数据的拟合程度对比

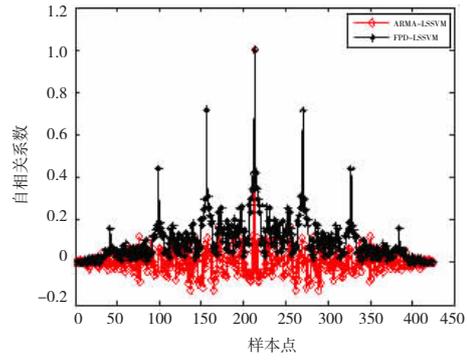


图 12 PDF-LSSVM 与 ARMA-LSSVM 训练模型残差的自相关分析

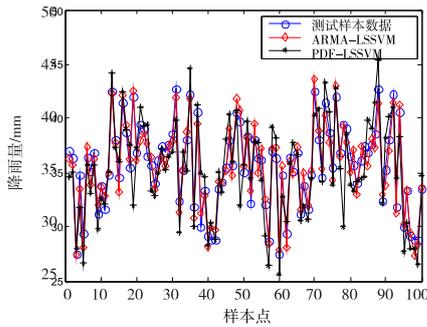


图 9 PDF-LSSVM 与 ARMA-LSSVM 模型预报结果与测试数据拟合程度对比

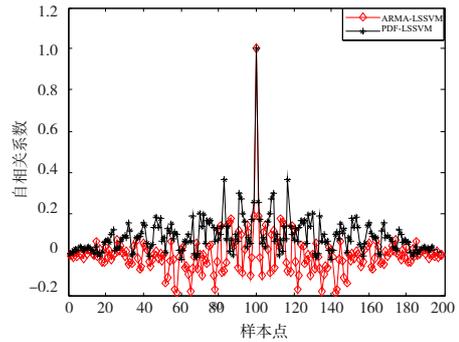


图 13 PDF-LSSVM 与 ARMA-LSSVM 预报残差的自相关分析

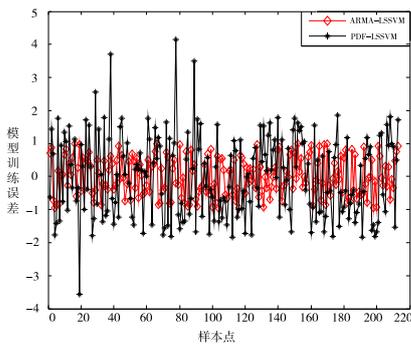


图 10 PDF-LSSVM 与 ARMA-LSSVM 模型训练残差

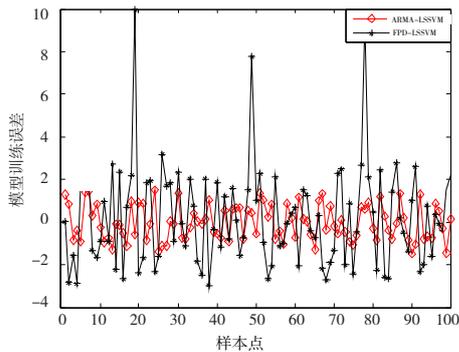


图 11 PDF-LSSVM 与 ARMA-LSSVM 模型预报残差

图 8 ~ 图 11 的对比结果表明,虽然两种模型训练的结果都能对训练样本数据进行拟合,两种模型训练精度都能满足要求,但是,与单纯的 PDF-LSSVM 模型相比,ARMA-LSSVM 的训练精度、预报精度、泛化性更高,ARMA-LSSVM 更具有实用价值。

图 12 和图 13 中,PDF-LSSVM 模型的训练残差与预报残差均不为白噪声,即 PDF-LSSVM 建模型没能提取出建模对象的全部信息,造成模型精度不高、泛化性差等问题。与之相比,ARMA-LSSVM 模型的训练残差与预报残差近似为白噪声,因此其模型结果具有更高精度与应用价值。

参考文献:

- [1] 孙翔,王景成.基于回归模型的城市长期水量预测[J].微型电脑应用,2010,26(11):7-9.
- [2] 才让加.化学数据的一元线性回归分析[J].青海师范大学学报:自然科学版,2005(2):13-15.
- [3] 姚伟.税收组合预测仿真研究[J].计算机仿真,2012,29(10):374-377.
- [4] 叶宗裕.非线性回归模型参数估计方法研究——以 C-D 生产函数为例[J].统计与信息论坛,2010,25(1):41-45.
- [5] 张金旺,刘红,华琳,等.非线性回归模型拟合生存资

- 料分析[J].数理医药学杂志,2009,22(6):641-642.
- [6] Ratkowsky D A. Nonlinear Regression Modeling: a unified practical approach[M]. New York: Marcel Dekker Inc., 1983.
- [7] 张新波. 时间序列模型在税收预测中的应用[J]. 湖南税务高等专科学校学报, 2010, 23(4): 30-32.
- [8] 林锦朗. 时间序列模型在海关税收预测中的应用[J]. 统计与咨询, 2009(1): 26-27.
- [9] 王时绘, 周健. 时间序列数学模型在税收分析中的应用[J]. 科技广场, 2011(7): 150-154.
- [10] 张伏生, 汪鸿, 韩悌, 等. 基于偏最小二乘回归分析的短期负荷预测[J]. 电网技术, 2003, 27(3): 36-40.
- [11] 孙智勇, 刘星. 税收增长预测的灰色理论模型研究[J]. 重庆大学学报: 社会科学版, 2010, 16(3): 41-45.
- [12] 郭晓君, 李大治, 褚海鸥, 等. 基于 GM(1,1) 改进模型的“两税”税收预测研究[J]. 统计与决策, 2014(4): 34-36.
- [13] 王敏. 税收收入预测方法的优选与应用[J]. 税务研究, 2009(10): 35-38.
- [14] Walczak B, Massart D L. Dealing with missing data: Part II[J]. Chemometrics and Intelligent Laboratory Systems, 2001, 58(1): 29-42.
- [15] Schafer J L, Graham J W. Missing data: Our view of the state of the art[J]. Psychological Methods, 2002, 7(2): 147-177.
- [16] Chen J, Bandoni A, Romagnoli J A. Outlier detection in process plant data[J]. Computers and Chemical Engineering, 1998, 22(4-5): 641-646.
- [17] 赵慧, 甘仲惟, 肖明. 多变量统计数据中异常值检验方法的探讨[J]. 华中师范大学学报: 自然科学版, 2003, 37(2): 133-137.
- [18] Victoria J H, Jim A. A survey of outlier detection methodologies[J]. Artificial Intelligence Review, 2004, 22(2): 85-126.
- [19] 成忠. PLSR 用于化学化工建模的几个关键问题的研究[D]. 杭州: 浙江大学, 2005.
- [20] 张玉, 尹腾飞. 支持向量机在税收预测中的应用研究[J]. 计算机仿真, 2011, 28(9): 357-360.
- [21] 刘碧森, 姚宇. 粗 SVM 理论及其在税收预测中的应用[J]. 仪器仪表学报, 2005, 26(8): 1530-1531.
- [22] 侯利强, 杨善林, 陈志强, 等. 基于遗传优化偏最小二乘支持向量机的税收预测研究[J]. 科技管理研究, 2014, 34(11): 197-200.
- [23] 肖苏, 熊焱. 基于灰度统计和神经网络的物流业税收预测模型[J]. 物流技术, 2013, 32(12): 131-134.
- [24] 傅俊, 朱莉. 基于残差控制的最小二乘支持向量机建模方法[J]. 计算机工程与应用(待发表).

## Research on Time Series Data Mining Based on Intelligent Integrated Particle Swarm Optimization Algorithm

ZHANG Jian

(College of Computer Science and Engineering, Sanjiang University, Nanjing 210012, China)

**Abstract:** An intelligent integrated architecture is proposed to address the problem that a single algorithm has the defect that can't dig all information in dealing with complex time series data. Four kinds of integration architecture have been given and their applications have been analyzed. Aiming at the time series data of a class of random noise interference, a series nested modeling structure is adopted, and the autoregressive moving average model of multiple double-population particle swarm optimization algorithm is proposed to dig the randomness trend in data. Meanwhile, The least squares support vector machine (LSSVM) based on probability density functions control (PDF) is proposed to dig the certainty trend in data, the parallel compensation of two models realizes the full excavation of data information. Through a set of experiments, the effect of proposed method is verified.

**Key words:** time series; Support Vector Machine (SVM); intelligent integrated; ARMA