

基于小波分析与支持向量机的时间序列预测

肖凡¹, 马捷中¹, 任岚昆²

(1. 西北工业大学 计算机学院, 陕西 西安 710072;

2. 西北工业大学 软件与微电子学院, 陕西 西安 710072)



摘要:时间序列广泛存在于工业、经济、军事等各个领域,时间序列预测是数据分析处理的一个重要方面。目前提出的预测模型大多基于“原始时间序列是无噪的”这一假定,而实际应用中,对时间序列去噪处理的好坏将直接影响预测的准确率,针对这一事实,使用小波分析对原始时间序列去噪。利用小波变换对时间序列进行多尺度分解,对各尺度上的细节序列使用阈值法去噪;使用支持向量机对重构后的各组小波系数进行预测并将结果融合,得到预测结果。实验结果表明,用于时间序列预测,能及时反应序列的变化趋势并具有较高的预测精度。

关键词:小波分析;多尺度分解;去噪;支持向量机;时间序列预测

中图分类号:TP18

文献标识码:A

文章编号:1671-654X(2011)06-0049-04

Time Series Prediction Based on Wavelet Analysis and Support Vector Machine

XIAO Fan¹, MA Jie-zhong¹, REN Lan-kun²

(1. School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China;

2. School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Time series is widespread in the industrial, economic, military fields and so on. Predicting the time series is one of the important aspects of data analysis and treatment. For the moment, most predicted models are based on the assumption that the original time series doesn't contain noise, but in the practical application, if the original time series couldn't be denoised properly, the accuracy of the prediction would be affected greatly. This paper uses wavelet analysis to denoise the original time series. Wavelet analysis could be utilized to analyze the time series in multiple scales and then the threshold method is used to denoise the detailed sequence in each scale; support vector machine(SVM) is applied to predict the reconstructed wavelet coefficient of each group and fusing all the predictions of them, the predicting results are got. Experimental results show that this method for time series prediction could timely response to the trend of time series and has high prediction accuracy.

Key words: wavelet analysis; multi-scale decomposition; denoising; support vector machine; time series prediction

引言

时间序列是一组有序的、随时间变化的数值序列,其中相邻数值间的时间间隔一般是相等的。世界上的许多事物、现象的发展变化都离不开时间,所以时间序列数据库的分布相当广泛。利用时间序列的相关性分析时间序列,预测随机机制在未来时刻的取值,是数据挖掘的一个重要分支。但目前提出的预测模型大多是基于“原始时间序列是无噪的”这一假定,而实际应用

中测得的时间序列常被各种形式、不同程度的噪声所污染。噪声产生冲突的规则,增加了系统的复杂性,弱化了序列的自相关程度,因而未经过去噪处理而建立的模型不能很好地刻画时间序列的内在规则,使得模型在实际预测中的泛化能力受到损害。

支持向量机(support vector machine, SVM)是 Vapnik 等提出的一类新的机器学习方法^[1]。由于其很快显现出出色的学习性能,该技术已成为机器学习界的研究热点,并在诸多领域都得到了成功的应用。支持

收稿日期:2011-07-22

修订日期:2011-09-06

基金项目:航空科学基金项目资助(2008ZD53035);陕西省自然科学基金项目资助(SJ08F20)

作者简介:肖凡(1987-),女,湖北天门人,硕士研究生,主要研究方向为测控与仿真、智能信息处理、故障预测与诊断。

向量机在高维小样本情况下有着很好的泛化能力,这个特性非常适用于时间序列预测的研究。同时,考虑到原始时间序列含有噪声将对预测精度产生较大影响,而小波分析是近年来迅速发展的一种很好的信号处理手段,具有良好的时频局域化特性,通过伸缩和平移对信号进行多分辨分析,能够聚焦到对象的任意细节^[2]。首先采用小波变换将时间序列进行多尺度分解,利用小波分解的多尺度特性提取原始序列中的高频信息和低频信息,由于噪声主要集中在高频部分,所以我们对高频系数使用小波阈值法^[3]进行去噪处理,再对去噪后的各子序列分别利用 SVM 进行回归预测,建立多层次、多尺度预测模型。

1 小波分析

1.1 小波变换原理

从信号处理的角度讲,小波变换是强有力的时频分析工具,是在克服傅里叶变换缺点的基础上发展而来的。平方可积函数 $f(t) \in L^2(R)$ 的连续小波变换定义为:

$$WT_f(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t) \cdot \psi\left(\frac{t-b}{a}\right) dt \quad (1)$$

$$= \langle f(t), \psi_{ab}(t) \rangle$$

其中,小波变换的核函数 $\psi_{ab} = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right)$ 是母小波 $\psi(t)$ 的时间平移 b 和尺度伸缩 a 的结果。

本文首先利用小波变换对原始时间序列进行多尺度分解,得到时间序列的近似系数和细节系数。这里采用 Daubechies (db) 紧支小波作为分解用到的小波基。给定时间序列 $X = \{x(t_i)\}$, 利用 db 小波对 X 进行多尺度分解,得到近似系数为 a_j , 各细节系数分别为 d_1, d_2, \dots, d_j (j 为最大分解层数)。各层细节序列和近似序列是原始序列 X 在相邻的不同频段上的分量。

1.2 小波阈值法去噪

应用小波变换去噪主要是利用其作为“数学显微镜”能较好地聚焦信号的局部结构特征,以及信号局部结构特征下所表现的奇异性相异于噪声所表现的奇异性,相对来说,信号的小波系数值必须大于那些能量分散且幅值较小的噪声的小波系数值。选择一个合适的阈值,对小波系数进行阈值处理,就可以达到去除噪声而保留有用信号的目的。该方法能得到原始信号的近似最优估计,并且具有非常广泛的适应性。对阈值 t 的选取,采用 Donoho 提出的通用阈值算法^[4],通用阈值由下式定义:

$$t = \sqrt{2 \log_2 s} \quad (2)$$

其中,对于小波包变换计算, $S = N \ln N$; 对于离散小波

变换计算, $S = N$ 。

Donoho 还提出了软硬阈值法:

$$\text{硬阈值法 } \hat{x} = T_h(Y, t) = \begin{cases} Y & |Y| \geq t \\ 0 & |Y| < t \end{cases} \quad (3)$$

$$\text{软阈值法 } \hat{x} = T_h(Y, t) = \begin{cases} \text{sgn}(Y)(|Y| - t) & |Y| \geq t \\ 0 & |Y| < t \end{cases} \quad (4)$$

其中 Y 是小波系数, \hat{x} 为小波阈值。

2 支持向量回归

2.1 支持向量回归算法

用 SVM 来估计回归函数,其基本思想就是通过一个非线性映射 ϕ , 将输入空间的数据 x 映射到高维特征空间 G 中,并在这个空间进行线性回归^[5]。设给定的样本数据 $\{x_i, y_i\}$, $x_i \in R^m, y_i \in R, i = 1, 2, \dots, s$ 。其中 y_i 为期望值, s 为数据点的总点数。一般采用下式来估计函数:

$$y = f(x) = (\omega \cdot \phi(x)) + b, \phi: R^m \rightarrow G, \omega \in G \quad (5)$$

对优化目标取值:

$$\min Q = \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\xi_i^* + \xi_i)$$

$$\text{s. t. } \begin{cases} y_i - (\omega \cdot \phi(x_i)) - b \leq \varepsilon + \xi_i^* \\ (\omega \cdot \phi(x_i)) + b - y_i = \varepsilon + \xi_i \\ \xi_i^*, \xi_i \geq 0 \quad i = 1, 2, \dots, s \end{cases} \quad (6)$$

式中: C 作为惩罚因子实现经验风险和置信范围的平衡折中; ξ_i, ξ_i^* 为松弛因子; ε 为损失函数。SVM 通过引入损失函数来解决回归问题,损失函数能够用稀疏数据点表示决策函数。不灵敏损失函数 ε 定义为:

$$L_i(y) = \begin{cases} 0 & |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon & |f(x) - y| \geq \varepsilon \end{cases} \quad (7)$$

引入拉格朗日乘子 a_i 和 a_i^* 把凸优化问题简化为最大化二次型:

$$\max_{a, a^*} W(a, a^*) = \sum_{i=1}^n y_i (a_i - a_i^*) - \varepsilon \sum_{i=1}^n (a_i + a_i^*) - \frac{1}{2} \sum_{i,j=1}^n y_i (a_i - a_i^*) (a_j - a_j^*) (x_i \cdot x_j)$$

$$\text{s. t. } \begin{cases} \sum_{i=1}^n a_i = \sum_{i=1}^n a_i^* \\ 0 \leq a_i \leq C \quad i = 1, 2, \dots, n \\ 0 \leq a_i^* \leq C \end{cases} \quad (8)$$

其中: C 用于控制模型的复杂度和逼近误差的平衡折中,越大则对数据的拟合程度越高; ε 用于控制回归逼近误差和模型的泛化能力。

对于非线性问题,用核函数来代替内积计算。常用的核函数有:多项式核函数、线性核函数、高斯径向

基核函数等。此时,回归函数表示为:

$$f(x) = \sum_{i=1}^n (a_i - a_i^*) K(x_i, x_j) + b \quad (9)$$

本文选用高斯径向基核函数(RBF),即:

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / \sigma^2) \quad (10)$$

2.2 支持向量机预测模型

对于时间序列 $\{x_1, x_2, \dots, x_n\}, i=1, 2, \dots, n, \{x_n\}$ 是目标预测值。建立自相关输入 $x_n = \{x_{n-1}, x_{n-2}, x_{n-3}, \dots, x_{n-m}\}$ 与输出 $y_n = \{x_n\}$ 之间的映射关系: $R^m \rightarrow R, m$ 为嵌入维数,可得用于支持向量机学习的样本:

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_m \\ x_2 & x_3 & \dots & x_{m+1} \\ \dots & \dots & \dots & \dots \\ x_{n-m} & x_{n-m+1} & \dots & x_{n-1} \end{bmatrix}, Y = \begin{bmatrix} x_{m+1} \\ x_{m+2} \\ \dots \\ x_n \end{bmatrix} \quad (11)$$

训练支持向量机的回归函数为:

$$y_t = \sum_{i=1}^{n-m} (a_i - a_i^*) K(x_i, x_t) + b, t = m+1, m+2, \dots, n \quad (12)$$

可得到第一步预测为:

$$\hat{x}_{n+1} = \sum_{i=1}^{n-m} (a_i - a_i^*) K(x_i, x_{n-m+1}) + b \quad (13)$$

其中, $x_{n-m+1} = \{x_{n-m+1}, x_{n-m+2}, \dots, x_n\}$, 得到一个样本 $x_{n-m+2} = \{x_{n-m+2}, x_{n-m+3}, \dots, x_n, \hat{x}_{n+1}\}$, 则第二步预测为:

$$\hat{x}_{n+2} = \sum_{i=1}^{n-m} (a_i - a_i^*) K(x_i, x_{n-m+2}) + b \quad (14)$$

一般可以得到第 i 步预测:

$$\hat{x}_{n+l} = \sum_{i=1}^{n-m} (a_i - a_i^*) K(x_i, x_{n-m+l}) + b \quad (15)$$

式中 $x_{n-m+l} = \{x_{n-m+l}, \dots, \hat{x}_{n+1}, \dots, \hat{x}_{n+l-1}\}$ 。

3 基于小波分析与支持向量机的预测模型

3.1 构建预测模型

结合小波分析与支持向量机的时间序列预测模型结构如图 1 所示,具体过程如下:

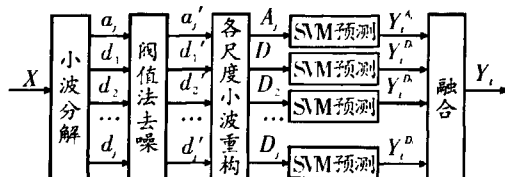


图1 预测模型结构图

1) 选择合适的小波基函数和小波分解层数 j , 将原始时间序列进行小波分解至 j 层, 得到相应的小波分解系数;

2) 对细节序列进行阈值去噪;

3) 重构各尺度下的小波系数;

4) 对重构后的小波系数分别建立支持向量机模型进行预测;

5) 将各预测结果进行融合, 得到最终预测值。

3.2 预测模型的评价标准

在评价预测结果时通常采用以下两项指标^[6]:

1) 平均绝对百分比误差(MAPE, 以%表示):

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{|y_i|}, y_i \neq 0 \quad (16)$$

2) 均方根误差(MSE):

$$MSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (17)$$

其中, y_i 为观测值, \hat{y}_i 为预测值, n 为预测个数。

3.3 预测模型的参数选择

本文主要采用平均绝对百分比误差 MAPE 作为评价标准来选择预测模型的参数。主要需要考虑的参数有: 高斯核参数 σ 、时间序列嵌入维数 m 、惩罚因子 C 、回归逼近误差控制参数 ε 。嵌入维数采用最小预测误差准则确定, 以达到较好的预测效果; ε 在预测精度要求范围内选取最大值, 惩罚因子 C 根据拟合误差和预测误差曲线变化规律进行调整, 以达到近优值, 从而提高预测模型的泛化能力。

4 数据仿真实验

4.1 时间序列实例预测

下面以某仪器测得的一组工业过程粘度数据为例, 验证上述基于小波分析与支持向量机的时间序列预测模型的实用性和有效性。对这组 200 个数据, 选取前 150 个用于训练, 后 50 个用于预测。

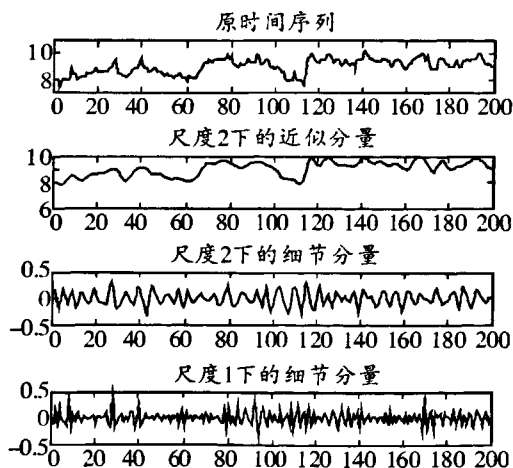


图2 原时间序列及各分量

对于小波分解采用的小波基函数和分解的级数, 应根据数据变化的具体情况和数据采样率进行恰当选

择。实验研究表明,当信号波动性较强而数据采样点过稀时,过高的小波分解级数会对信号确定性的波动现象进行过滤,引起对原信号恢复的失真。经过实验比较,对本实验时间序列小波分解选取 db6 小波基,分解级数为 2,可在尽量保持时间序列数据波动特征的情况下去除噪声。原时间序列及其分解后的各个分量如图 2 所示。

对各高频分量进行阈值法去噪,得到去噪后的各分量序列。重构各分量,得到去噪后的时间序列,如图 3 所示,可以看出大部分噪声被消除。

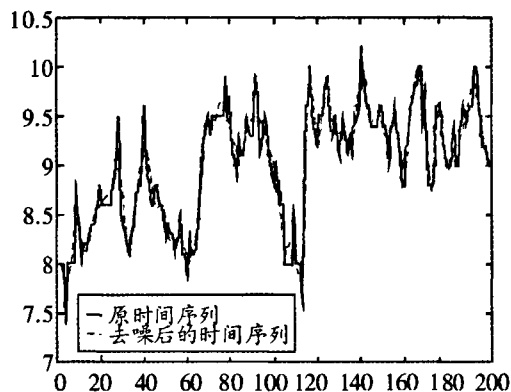


图 3 原时间序列与去噪后的时间序列

去噪后的各分量序列都有 200 个数据,前 150 个作为训练集,后 50 个作为测试集,分别选择合适的支持向量机进行预测,得到各分量序列的预测值,融合各分量预测值就得到最终时间序列预测值。

以近似分量为例,选取嵌入维数 $m=5$, 高斯径向基核参数 $\sigma=2$, 惩罚因子 $C=10$, 回归逼近误差控制参数 $\varepsilon=0.001$, 预测结果如图 4 所示。

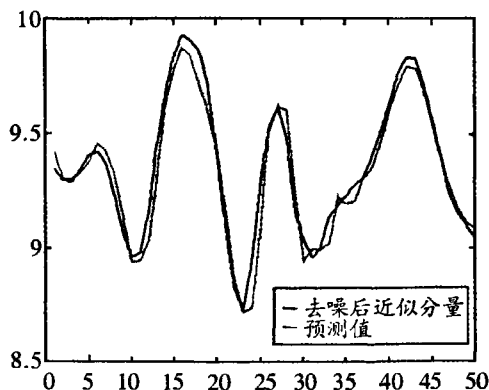


图 4 近似分量的预测结果

各分量预测值融合,所得预测结果如图 5 所示。

4.2 预测结果分析

使用单一支持向量机对这组时间序列进行预测,可得如图 6 所示预测结果。

由图 6 可看出,未去噪的时间序列直接进行支持向量机预测时,对数据序列的变化趋势反应较慢,有一定延迟,而图 5 所示结合小波分析的支持向量机能及时反应时间序列的变化趋势。

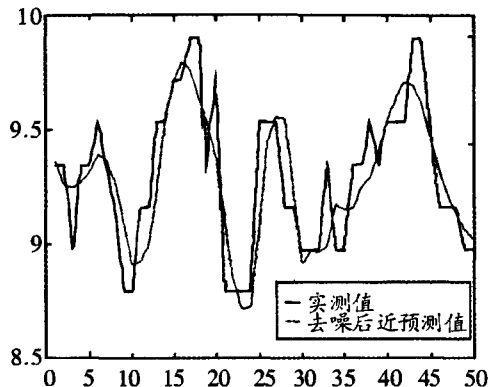


图 5 去噪后的预测结果与实测值对比

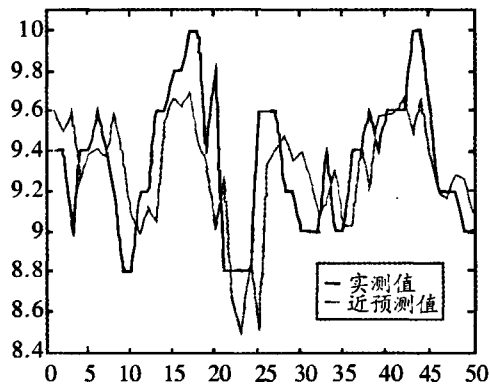


图 6 未去噪时间序列 SVM 预测

两种方法预测时间序列的和如下表所示。

两种方法的预测结果表		
预测方法	MAPE/%	MSE
单一 SVM	2.62	0.328 3
本文 (Wavelet + SVM)	1.63	0.196 3

由上表可看出,基于小波分析与支持向量机的时间序列预测模型在预测精度上较单一支持向量机也有显著提高,表明该方法用于时间序列预测的有效性。

5 结束语

针对实际应用中测得的时间序列通常含有噪声的情况,本文建立了一个结合小波分析与支持向量机的时间序列预测模型。利用小波变换对时间序列进行多尺度分解,对各高频细节分量使用阈值法去噪;使用支持向量机对重构后的各小波系数进行预测并将结果融

(下转第 57 页)

1. 25° 。约束条件为 $0.073 \leq S \leq 0.083$, $8.0 \leq \beta_{trailing} \leq 30.0$, $C_l \geq 0.83$ 。 S 为翼型截面积, $\beta_{trailing}$ 为翼型后缘角, C_l 为升力系数。适应度函数为:

$$f = 1/C_d \quad (6)$$

面积和后缘角约束罚函数同(4),(5)式,升力系数约束罚函数为:

$$F_{Cl} = 10^4 \times (C_l - 0.83) - 10^4 \quad (7)$$

表 4 阻力系数优化结果与参考翼型对比

模型	C_l	C_d	S	$\beta_{trailing}$
参考翼型	0.832	0.050	0.078	8.676
优化翼型	0.860	0.015	0.074	24.59

表 5 设计变量约束范围及优化结果

设计变量	约束范围	优化结果
R_{te}	$0.001 \leq R_{te} \leq 0.1$	0.004 5
β/rad	$0.001 \leq \beta \leq 0.6$	0.407 5
β_{te}/rad	$0.001 \leq \beta_{te} \leq 0.6$	0.401 5
β/rad	$0.001 \leq \beta \leq 0.6$	0.407 5
$\Delta z_{te}/c$	$-0.001 \leq \Delta z_{te}/c \leq 0.001$	-0.000 397
b_1	$0.0 \leq b_1 \leq 0.5$	0.066
b_2	$0.0 \leq b_2 \leq 0.5$	0.150
b_3	$0.0 \leq b_3 \leq 0.5$	0.014
b'_1	$0.0 \leq b'_1 \leq 0.5$	0.201
b'_2	$0.0 \leq b'_2 \leq 0.5$	0.232
b'_3	$0.0 \leq b'_3 \leq 0.5$	0.165

优化结果如图 12 ~ 图 14 所示,表 4 给出了优化结果与参考翼型的对比,设计变量约束范围及优化结果见表 5。由表 4 可知,基于本算例给定的约束条件,优化后的翼型较参考翼型阻力系数有了大幅减小(70%),展示了本文方法的有效性。

5 结束语

CST 参数化方法直接采用翼型几何特征作为控制

参数,相比常见参数曲线翼型表达方法更能反映翼型特有的气动敏感性,有利于遗传算法搜索寻优。本文在翼型优化中引入了 CST 参数化方法,发展了基于 CST 参数化方法的气动优化遗传算法,并对其进行了算例考核,计算结果展示了本文方法用于翼型气动优化的效果。由优化结果来看,将 CST 参数化方法与遗传算法相结合,用于翼型气动优化问题是可行的。

参考文献:

- [1] 周明,孙树栋. 遗传算法原理及应用[M]. 北京:国防工业出版社,1999:39-40.
- [2] Brenda M Kulfan, John E Bussoletti. Fundamental Parametric Geometry Representations for Aircraft Component Shapes [A]. AIAA 2006-6948.
- [3] 孙明华,崔海涛,温卫东. 基于精英保留遗传算法的连续结构多约束拓扑优化[J]. 航空动力学报,2006,21(4): 732-737.
- [4] Blazek J. Computational Fluid Dynamics: Principles and Applications [M]. Oxford: Elsevier Science, 2001: 1-215.
- [5] 刘学强. 基于混合网格和多重网格上的 N-S 方程求解及应用研究[D]. 南京:南京航空航天大学,2001.
- [6] Frederic J Blom. Considerations on the Spring Analogy [J]. International Journal for Numerical Methods in Fluids, 2000, 32: 647-668.
- [7] 王小平,曹立明. 遗传算法—理论、应用与软件实现[M]. 西安:西安交通大学出版社,2002:1-122.
- [8] Marco C, Macelo H, Ernani V. A Study of the CST Parameterization Characteristics [A]. AIAA 2009-3767.
- [9] 周春华,伍贻兆. 非结构网格上 Euler 方程有限体积解法的改进[J]. 南京航空航天大学学报,2003,35(3): 313-317.

(上接第 52 页)

合,得到预测结果。通过实例分析,本方法与单一支持向量机对时间序列预测的结果相比,能及时反应序列的变化趋势并具有较高的预测精度,有较好的预测效果。

参考文献:

- [1] Vapnik V. The Nature of Statistical Learning Theory [M]. New York: Springer-Verlag, 1995.
- [2] 文莉,刘正士,葛运建. 小波去噪的几种方法[J]. 合肥工业大学学报,2002(2): 167-172.

- [3] Donoho D L. Adapting to Unknown Smoothness Via Wavelet Shrinkage [J]. Journal of American Stat Assoc, 1995, 90: 1200-1224.
- [4] Donoho D L. Denoising by Soft-thresholding [J]. IEEE Transaction on Information Theory, 1995(3): 613-627.
- [5] 杨金芳,翟永杰,王东风,等. 基于支持向量回归的时间序列预测[J]. 中国电机工程学报,2005(17): 110-113.
- [6] 李满,徐进军. 基于小波分析与 LSSVM 的滑坡变形预测[J]. 大地测量与地球动力学,2009(4): 127-130.

基于小波分析与支持向量机的时间序列预测


作者：

[肖凡](#)，[马捷中](#)，[任岚昆](#)，[XIAO Fan](#)，[MA Jie-zhong](#)，[REN Lan-kun](#)

作者单位：

[肖凡, 马捷中, XIAO Fan, MA Jie-zhong\(西北工业大学计算机学院, 陕西西安, 710072\)](#)，[任岚昆, REN Lan-kun\(西北工业大学软件与微电子学院, 陕西西安, 710072\)](#)

刊名：

[航空计算技术](#)

英文刊名：

[Aeronautical Computing Technique](#)

年，卷(期)：

2011, 41(6)

本文链接：http://d.g.wanfangdata.com.cn/Periodical_hkjsjs201106013.aspx