改进嵌入维数和时间延迟计算的GP预测算法

吕 威[1,2,3] 王和勇[4] 姚正安[1,2] 李 磊[1]

(中山大学软件研究所 广州 510080)[1] (中山大学数计学院 广州 510080)[2]

(北京师范大学珠海分校信软学院 珠海 519085)[3] (华南理工大学电子商务学院 广州 510275)[4]

摘 要 改进了混沌系统中的两个重要特征量:嵌入维数和时间延迟的计算,根据计算得出的上述两个参数重构相空间;然后在相空间中作轨迹的线性拟合,选择轨迹中的最近邻点作一次性的预测。提出的算法在相空间中很好地把轨迹的线性拟合与最近邻方法结合起来,解决了现有的时间序列分析和预测算法中主观性太强的缺点,通过对话务量时间序列和太阳黑子时间序列的验证,与其它算法相比,该算法的分析结果稳定而准确、预测精度高、运行时间比较短。

关键词 嵌入维数,时间延迟 时间序列,分形,最近邻预测

GP Predicate Algorithm Based on the Improved Computing of Embedding Dimension and Time Delay

LU Wei[1,2,3] WANG He-yong[4] YAO Zheng-an[1,2] LI Lei[1]

(Software Research Institute of SUN YAT-SEN University, Guangzhou 510080, China)[1]

(Mathematics Department of SUN YAT-SEN University, Guangzhou 510080, China)[2]

(School of Information Technology and Sofeware Engineering of Beijing Normal University Zhuhai Campus, Zhuhai 519085, China)[3]

(College of E-Business of South China University of Technology, Guangzhou 510275, China)[4]

Abstract This paper improved the computing of two important characteristic measure embedding dimension and time delay in chaos system, and reconstructed phase space based on these two parameter. And then it simulated track linearly in phase space, selecting nearest neighbor for one time predicate. The new algorithm combines the linear track simulation and the nearest neighbor method well, sloving disadvantage that the subjectivity is too strong in existing time serials and predicate method. By validating the phone number time serials and sunspot serials, comparing other algorithm, the analysis result of our method is steady and exact, predicate precision of it is high and the running time of it is short.

Keywords Embedding dimension, Time delay, Time serials, Fractal, Nearest neighbor predicate

1 引言

分形(fractal)是由 Benoit Mandelbrot 在 1975 年发表的奠基性论文中,为高度不规则的集合的命名,含有破碎的和不规则的两重含义[1]。离散非线性系统常表现为时间序列的形式,因此借助分形理论来研究时间序列分析和预测问题是非常有益的。

重构相空间中的 Takens 定理[1]表明,可以找到一个合适的嵌入维,在这个嵌入维空间里可以把有规律的轨迹(吸引子)恢复出来。因此,可以把时间序列转化到合适的嵌入维空间中研究。

Grassberger 和 Procaccia 于 1984 年提出 GP 算法[2],用于从时间序列中提取信息,Grassberger 和 Procaccia 还给出了关联维数 d、信息维数 σ 和豪斯道夫维数 D 之间的关系:$d \leqslant \sigma \leqslant D$。

分形理论有着广泛的应用,如文献[3]分析了分形理论在地震学中的应用,文献[4] 利用分形理论对 Lennard-Jones12-6 流体气液界面性质进行了研究,文献[5]提出可以用分形理论来分析网络流量。GP 算法也可用于预测中,如文献[6] 用分形理论来预测经济流通领域的危机点,文献[7] 研究了时空复杂问题中的多分形预测问题。

本文提出将 GP 算法用于时间序列分析和预测,很好地解决了一般预测算法中人工干预的主观性问题。本文的主要创新在于改进了 GP 算法中嵌入维数和时间延迟的计算,将分形理论和最近邻算法有机结合在一起,形成了一种新的预测算法,降低了一般预测算法中需人工干预来选取模型参数的主观影响,跟相似的算法相比,在精度上有了较大的提高。本文第 2 节介绍 GP 算法,第 3 节提出了改进嵌入维数和时间延迟计算的 GP 预测算法,第 4 节用试验来验证我们的算法,最后给出了结论和进一步研究的方向。

2 GP 算法

用分形理论研究时间序列问题首先要考虑的是相空间重构问题,就是将时间序列 $\{x_i\}_{i=1}^{n}$ 转化为:

$$\{X_i(m,\tau) \mid X_i(m,\tau)=(x_i, x_{i+\tau}, \cdots, x_{i+(m-1)\tau}), \quad 1 \leqslant i \leqslant n-(m-1)\tau\} \tag{1}$$

到稿日期:2008-06-24 本文受国家自然科学基金项目(No. 10531040)资助。

吕 威 男,博士生,讲师,主要研究数据分析、机器学习,E-mail:luwei00@126.com。

其中 m 为嵌入的维数，τ 为时间延迟，是采样时间间隔 Δt 的整数倍。经过重构相空间，时间序列从一维空间重构成 m 维欧氏空间。

Grassberger 和 Procaccia 提出的 GP 算法可以从时间序列中提取信息。

GP 算法流程：

Step 1　重构相空间：按式(1)把长度为 n 的一维序列重构成 $N_m=n-(m-1)\tau$ 个 m 维时间序列。

Step 2　计算关联积分函数：

$$C_m(r)=\frac{2}{N_m(N_m-1)}\sum_{i=1}^{N_m}\sum_{j<i}H(r-r_{ij})$$

其中，H 为 Heavicide 函数

$$H(x)=\begin{cases}1 & x>0\\ 0 & x\leqslant 0\end{cases}$$

$$r_{ij}=d(X_i,X_j)=\left[\sum_{l=0}^{m-1}(x_{i+l\tau}-x_{j+l\tau})^2\right]^{\frac{1}{2}}$$

Step 3　估计关联维数：当 r 取充分小时，关联积分函数逼近下式：

$$\ln C_m(r)=\ln C+D(m)\ln r$$

因此，m 维空间数据的关联维数

$$D(m)=\lim_{r\to 0}\frac{\partial \ln C_m(r)}{\partial \ln r}$$

Step 4　估计维数：把 $D(m)$ 与以前计算出的 $D(i)$，$(i<m)$ 比较，若 $D(m)$ 不随嵌入维数 m 的升高而改变，则

$$D_2=\lim_{r\to 0}D(m)$$

就是该系统的关联维数。否则增加嵌入维数，转 Step 1。

重构相空间的预测算法对时间序列预测具有较强的确定性，且在必要嵌入维较低的情况下，通常表现出令人满意的预测结果。但它的预测精度在某个嵌入维达到最大值后，一般地将随嵌入维的升高而下降[2]。文献[8]认为原因在于重构集的全局指数谱的变化。

3　改进嵌入维数和时间延迟计算的 GP 预测算法

GP 算法中的相空间重构有着多种改进方法[9,10]。对嵌入维数 m 的确定，文献[11]提出了虚假最近邻点法(False Nearest Neighbors，FNN)，文献[12]提出了奇异值分解法(Singular Value Decomposition，SVD)；而对时间延迟 τ 的计算，一般有自相关函数法、互信息法[13]、平均位移法(Average Displacement，AD)[14]等。为了适合将 GP 算法用于时间序列分析和预测，我们改进了嵌入维数 m 和时间延迟 τ 的计算。

3.1　时间延迟和嵌入维数的改进计算

在利用 GP 算法提取时间序列特征的过程中，复自相关法和 GP 算法会形成循环，如图 1 所示。

图 1　GP 算法和复自相关法所形成的循环

本文改进了上述方法，提出使用自相关法作为起始点，从时间序列直接计算出适合二阶(嵌入维数 $m=2$)情况的时间延迟 $\tau 0$，将 $\tau 0$ 应用于上述循环来优化嵌入维数 m 和时间延迟 τ，其流程如图 2 所示。

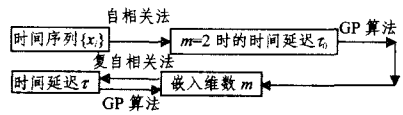

图 2　以自相关算法为入口的嵌入维数和时间延迟的计算流程

在实际运算中，发现通过 GP 算法计算出的关联维数 $D(m)$ 与嵌入维数 m 之间的关系如图 3 所示。随着 m 的增大，$D(m)$ 出现一波又一波的升浪，每一波末端的 $D(m)$ 值相近，都可以作为关联维数 $D2$ 的估计值。对某一个时间延迟 τ，若计算出的升浪越长，由该对 (τ,m) 值所重构的相空间中轨迹的维数就越稳定，这样的相空间也更适合于恢复系统的运行轨迹。

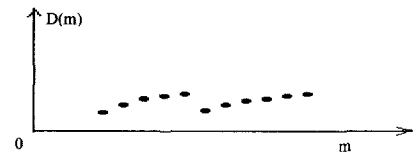

图 3　关联维数 $D(m)$ 随嵌入维数 m 变化的示意图

复自相关法和 GP 算法的迭代循环应该反复进行，直到第一波升浪的长度取得极大值为止；此时的时间延迟 τ 和第一波升浪长度 m 就是时间延迟和嵌入维数的合适的值。

综合上述，提出的改进嵌入维数和时间延迟的计算流程如下：

算法 1　时间延迟和嵌入维数的改进计算

Input 时间序列 $\{x_t\}_{t=1}^{n}$

Step 1　用自相关法计算时间延迟 τ。

Step 2　根据 $\tau 0$ 用 GP 算法计算合适的嵌入维数 m、关联维数 d。

Step 3　$\tau 0=\tau, m0=m, d0=d$。

Step 4　根据 m 用复自相关法计算适合 m 阶情况的时间延迟 τ。

Step 5　根据 τ 用 GP 算法计算合适的嵌入维数 m、关联维数 d。

Step 6　若 $m>m0$，则转 Step 3，否则退出。

Output $\tau 0, m0, d0$。

上述算法是通过循环迭代不断优化所得的参数，从而计算出能使相空间中轨迹的维数相对稳定的时间延迟和嵌入维数的值。

3.2　改进的 GP 一次性预测算法

混沌时间序列的预测所遇到的最大困难是，在映射 $x_{n+1}=f(x_n)$ 的作用下邻点可能会变成分离点，使得若要根据近邻点来拟合局部的 f 变得非常困难。考虑到混沌系统的这种分离性，本文采用最近邻算法，即在轨迹中选取与最末一点欧氏距离最小的一点，把它未来的值作为预测值。

设轨迹的最后一点是 X_N，$\{X_i\}_{i=1}^{N}$ 中 X_i 的坐标是 $(a_{i1}, a_{i2},\cdots,a_{im})$，则相空间中 X_NX_i 之间的距离最小的 θ_i 值应是

$$\hat{\theta}_i=\begin{cases}0 & \text{if } \theta_i<0\\ \theta_i & \text{if } 0\leqslant\theta_i<1\\ undefined & \text{if } \theta_i\geqslant 1\end{cases}\qquad(2)$$

其中

$$\theta_i=\sum_{j=1}^{m}(a_{i+1,j}-a_{ij})(a_{Nj}-a_{ij})/\sum_{j=1}^{m}(a_{i+1,j}-a_{ij})^2\qquad(3)$$

X_N 与线段 $X_iX_{i+1}(i=1,2,\cdots,N-2)$ 上最短距离是：

$$D_i=\sqrt{\sum_{j=1}^{m}(z_{ij}-a_{Nj})^2}$$
$$=\sqrt{\sum_{j=1}^{m}((1-\theta_i)a_{ij}+\theta_i a_{i+1,j}-a_{Nj})^2}$$
$$=\sqrt{\sum_{j=1}^{m}((a_{i+1,j}-a_{ij})\theta_i+a_{ij}-a_{Nj})^2} \quad (4)$$

算法 2 求 X_N 到折线段 $X_1X_2\cdots X_{N-1}$ 的最近点的算法

Input 相空间中的点集 $\{X_i\}_{i=1}^{N}$

Step 1 初始化：$i=1$，$\min D=+\infty$；

Step 2 按式(2)(3)和式(4)计算 $\hat{\theta}_i$ 和 D_i；

Step 3 若 $\min D<D_i$，则记下 $\min D=D_i$，$\min I=i$，$\min\theta=\hat{\theta}_i$；

Step 4 $i=i+1$。若 $i<N-1$，转 Step 2，否则退出；

Output $\min I$，$\min\theta$。

如图 4 所示，设 X_N 的最近邻点 $Z_N=(1-\theta_i)X_i+\theta_iX_{i+1}$，经过时间 k(k 为正整数)该点到达

$$Z_{N+k}=(1-\theta_i)X_{i+k}+\theta_iX_{i+k+1}$$

此时，把 Z_{N+k} 的最后一维作为时间序列 $\{x_i\}_{i=1}^{n}$ 的后 k 步的预测值：

$$x_{n+k}=(1-\theta_i)a_{i+k,m}+\theta_i a_{i+k+1,m} \quad (5)$$

特别地，如果时间序列的数据都是连续的，中间没有断点，则有

$$a_{i+k,m}=x_{i+(m-1)\tau+k},a_{i+k+1,m}=x_{i+(m-1)\tau+k+1}$$

此时预测值

$$x_{n+k}=(1-\theta_i)x_{i+(m-1)\tau+k}+\theta_i x_{i+(m-1)\tau+k+1}$$

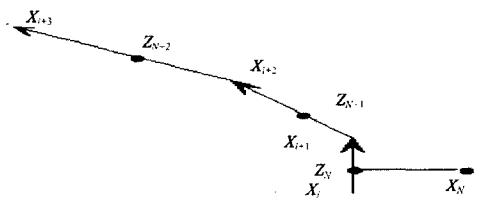

图 4 一次性预测方法示意图

采用这种一次性的预测方法，预测的步长增大了，但可以避免误差的积累。

提出的改进嵌入维数和时间延迟计算的 GP 预测算法可以最终描述为：

算法 3 改进嵌入维数和时间延迟计算的 GP 预测算法

Input 时间序列 $\{x_i\}_{i=1}^{n}$，预测长度 l

Step 1 按算法 1 计算时间延迟 τ，嵌入维数为 m；

Step 2 重构相空间：按式(1)把长度为 n 的一维序列重构成 $Nm=n-(m-1)\tau$ 个 m 维时间序列；

Step 3 寻找最近邻点：按算法 2 计算最近邻点的〈i，θ〉；

Step 4 从最近邻点推断：for $k=1,2,\cdots,l$：按式(5)计算出 x_{n+k}。

Output x_{n+1}，x_{n+2}，$\cdots x_{n+k}$

由此可见，在本文提出的改进嵌入维数和时间延迟计算的 GP 预测算法中，分形理论和最近邻算法创造性地、有机地结合在一起，形成一个完备的、无需人工干预的算法。

4 试验结果

我们选取了两个实际问题中的时间序列——话务量时间序列和太阳黑子时间序列来验证文中提出的算法的效果。文中还选取了最常用的 AR 算法来与线性轨迹最近邻预测算法进行结果对比。本文的实验都在 PC 机上进行，CPU 是赛扬 1.7G，内存为 256M DDR；操作系统为 Windows 2000 Server，在 Microsoft Visual C++ 6.0 上开发，由 Matlab 7.0 的数学库提供 AR 算法的实现。

4.1 话务量预测

在以下的实验中也将沿用上述的含义和处理方法。最后一列是我们方法的结果。结果见表 1 和图 5。

表 1 平日话务量预测结果

Time(时间)	Actual(实际值)	AR(24)	AR(100)	FractalNT	Fractal
2000022920	372.8	334.9	226.4	336.3	357.5
2000022921	443.4	280.9	229.0	370.8	419.2
2000022922	258.6	223.1	171.9	336.3	262.4
2000022923	117.3	169.7	107.2	370.8	112.2
MSE		7965.8	18755.7	19225.6	215.0
MRE		26.31%	32.44%	68.08%	3.84%
Running Time(ms)		5	5	94	94

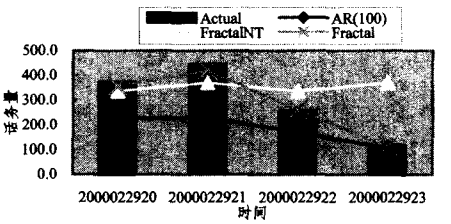

图 5 平日话务量预测结果图

从上述结果可以看出，改进嵌入维数和时间延迟计算的 GP 预测算法的预测精度远远高于其它算法，它不但反映了话务量时间序列的变化趋势，而且其预测数值与真实值也相当吻合。在运行时间方面，分形算法的运行时间虽然多于相对简单的 AR 算法，但也能在相当短的时间(<0.01 秒)内完成。

4.2 太阳黑子数据预测

以 1903 年 5 月到 2000 年 8 月共 1168 个月均太阳黑子数据为训练数据，预测 2000 年 9—12 月的月均太阳黑子数据。

时间分析算法计算出：时间延迟为 12，嵌入维数为 7，关联维数为 3.0647。预测结果见表 2 和图 6。从上述结果可以看出，在这次实验中，无论是平均平方误差还是平均相对误差，改进嵌入维数和时间延迟计算的 GP 预测算法的预测结果都大大优于 AR 算法和没有改进的最近邻算法，其精度分别提高了约 13.62%和 4.36%。随着样例的增加，分形算法的运行时间比前面的实验有轻微的增加。

表 2 2000 年 9—12 月月均太阳黑子数的预测结果

Time	Actual	AR(24)	AR(100)	FractalNT	Fractal
200009	109.7	129.3	124.5	135.8	110.6
200010	99.4	133.5	131.3	106.8	116.0
200011	106.8	132.2	127.9	120.0	103.4
200012	104.4	128.5	121.9	106.0	97.1
MSE		693.0	497.9	228.2	85.0
MRE		24.75%	20.54%	11.28%	6.92%
Running Time(ms)		5	6	3953	3953

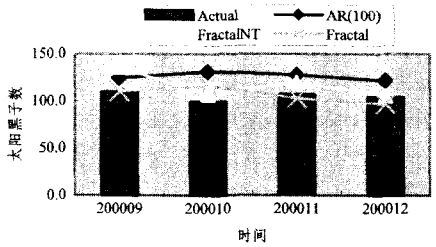

图 6　2000 年 9—12 月月均太阳黑子数的预测结果图

结束语　本文提出了改进嵌入维数和时间延迟计算的 GP 预测算法，既有比较严密的数学基础，又在实际应用中取得了良好的效果：分析结果稳定而准确、预测精度高、运行时间比较短，而且也不需人工干预。因此，它在时间序列研究的领域中具有重要的理论意义，进一步在与时间序列有关的众多应用领域中具有广阔的应用前景。

因为分形理论和近邻预测方法都是新兴的理论，还远远没有达到成熟的阶段，结合分形理论和近邻预测方法的时间序列研究方法还会有进一步的发展。具体来说，如嵌入维数和时间延迟的计算、轨迹的拟合、近似样例的选择、查询样例小邻域内函数的拟合等方面，都还存在改进的空间，值得作进一步的研究，以提出更好的理论和算法。

参 考 文 献

[1] 吕金虎，陆君安，陈士华. 混沌时间序列分析及其应用[M]. 武汉：武汉大学出版社，2002

[2] Grassberger P, Procaccia I. Dimension and Entropy of Strange Attractors from a Fluctuating Dynamic Approach[J]. Physica, 1984, 13D: 34-54

[3] 李信富，李小凡，武晔. 分形理论在地震学中的应用研究[J]. 地球物理学进展，2007(2)

[4] 王德明，曾丹苓，刘娟芳. 利用分形理论研究气液界面特性[J]. 工程热物理学报，2004(5)

[5] 杨新宇，曾明，赵瑞，等. 分形理论在网络流量分析中的应用综述[J]. 计算机工程，2004(23)

[6] Mansurov K. Forecasting currency crises by fractal analysis techniques[J]. Studies on Russian Economic Development, 2008, 19(1): 96-103

[7] Schertzera D, Lovejoy S. Space-time complexity and multifractal predictability[J]. Physica A: Statistical Mechanics and its Applications, 2004, 338: 173-186

[8] 候越先，可丕廉，王雷. 适用于高必要嵌入维的混沌时间序列预测算法[J]. 天津大学学报，1999, 32(5): 594-598

[9] 郑会永，刘华强，戴冠中. 时间序列分维的改进 GP 算法[J]. 西北工业大学学报. 1998, 16(1): 28-32

[10] Albano A M, Muench J, Schwartz C, et al. Singular-value decomposition and the Grassberger-Procaccia algorithm[J]. Phys. Rev. A, 1993, 38: 3017-3026

[11] Kennel, Mathew B, Brown R, et al. Determining embedding dimension for phase-space reconstruction using a geometrical construction [J]. Phy Rev A, 1992, 45: 3403-3411

[12] Broomhead D, King G. Extracting qualitative dynamics from experimental data [J]. Physica D, 1986, 20: 217-236

[13] Fraser A M, Swinney H I. Independent coordinates for strange attractors from mutual information [J]. Phys. Rev. A, 1986, 33: 1134-1140

[14] Rosenstein M T, Collins J J, Carlo D L J. Reconstruction expansion as a geometry-based framework for choosing proper delay times[J]. Physica D, 1994, 73: 82-98

(上接第 144 页)

结束语　本文在分析国内外的工作基础之上，针对已有的 Ctree 进行改进，提出了一种以结点路径相同为原则的索引结构 FC-Index，以对 XML 文档进行规范的结构处理，从而提高查询效率，针对合并后的结构，提出一种有效的查询算法，以对合并后的元素进行有效的查询。基于不同数据集的试验结果表明，本文提出的基于 FC-Index 的查询处理方法可以有效提高查询效率。

参 考 文 献

[1] XMARK(TheXML-benchmarkproject)[OL]. http://monetdb.cwi.nl/ xml

[2] Zou Qinghua, Liu Shaorong, Chu W W[C]// Proc. Ctree: A Compact Tree for indexing XML Data. Washington, DC, USA, 2004: 12-13

[3] Kaushik R, Bohannon P, Naughton J F, et al. Covering indexes for branching path queries[C]// SIGMOD. 2002

[4] Kaushik R, Bohannon P, Naughton J F, et al. Updates for structure indexes[C]// VLDB. 2002

[5] Kaushik R, Krishnamurthy R, Naughton J F, et al. On the integration of structure indexes and inverted lists[C]// SIGMOD. 2004

[6] Kaushik R, Shenoy P, Bohannon P, et al. Exploiting local similarity for efficient indexing of paths in graph structured data[C]// ICDE. 2002

[7] Dong Xin, Halevy A. Indexing Dataspaces[C]// SIGMOD. 2007

[8] Goldman R, Widom J. Enabling query formulation and optimization in semistructured databases[C]// On VeryLargeData Bases (VLDB). 1997

[9] Milo T, Suciu D. Index structures for path expressions[C]// ICDE. 1999

[10] Li Q, Moon B. Indexing and querying XML data for regular path expressions[C]// VLDB. 2001

[11] Kaushik R, Krishnamurthy R, Naughton J F, et al. On the Integration of Structure Indexes and Inverted Lists[C]// SIGMOD. 2004