科学数据时间序列的预测方法

周巧临，傅 彦

(1. 西南科技大学计算机学院 四川 绵阳 621010; 2. 电子科技大学计算机科学与工程学院 成都 610054)

【摘要】针对传统的时间序列分析方法预测科学数据效果较差的特点，提出了一种结合自组织神经网络和灰色理论的时间序列预测方法。该方法利用度量时间序列相似性距离函数，将时间序列按照其变化规律分成不同的类别，并在GM算法中针对白化参数进行优化，对科学数据时间序列进行自组织聚类，针对各类别采用灰色理论建立预测模型。试验表明，该模型适合科学数据的变化特点，提高了预测精度。

关 键 词 神经网络; 灰色理论; 时间序列; 预测
中图分类号 TP311; TP391 文献标识码 A

A Method of Time Series Forecasting for Scientific Data

ZHOU Qiao-lin，FU Yan

(1. School of Computer, Southwest University of Science Technology Mianyang Sichuan 621010;
2. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 610054)

Abstract Traditional methods have poor efficiency and effect to deal with the scientific data series forecasting. In this paper, a forecasting algorithm based on grey theory and self-organized map neural networks is proposed. Firstly, the scientific data time series cluster in self-organized mannar. Then the forecast model is established with grey theory. In clustering, a distance criterion is proposed to scale the difference between series. In grey theory, the whiten parameter is optimized. The experiments show that this algorithm surpasses those traditional forecasting methods in precision and time efficiency.

Key words neural network; grey theory; time series; forecasting

大规模科学计算是对连续变化的科学现象进行离散抽样的数值模拟，计算过程除需花费大量的计算资源外，还必然会产生大规模的数值模拟数据。仿真物理现象的数值模拟数据由数值计算程序通过一系列时间步产生，是对连续空间数据离散采样、带有网格结构的时间序列[1-2]。当数据规模增大时，时间序列的分析和处理都因为数据特征复杂、规模庞大而难以进行。数值模拟应用中，大规模科学数据的时间序列分析[3-5]，已经成为影响科学数据挖掘的一个重要瓶颈问题。

科学数据具有几何特性、拓扑结构、物理属性和时变性四个基本特征。特别是对于时变性来讲，存在对时间规则的网格数据和对时间步间隔变化的网格数据。常用的时间序列分析方法假定数据时间间隔相等，因此不适于科学数据时间序列分析。影响科学现象的物理因素是多种多样的，各种各样的物理因素对物理现象的具体影响仍需进一步研究，因此对科学数据时间序列的预测常具有灰色性特征。本文建立灰色预测模型，即以自组织神经网络聚类后的时间序列为基础，运用灰色模型预测科学数据时间序列的变化。

1 时间序列自组织聚类

自组织映射神经网络广泛地应用于将高维数据映射到低维空间进行分析，并且映射尽可能地保持输入数据之间的关联。

自组织特征映射是一个二维空间排列的人工神经网络，其中的每个神经元通过无监督学习过程进行调整，使之对应于特定的不同输入信号模式或模式分类。从根本上而言，在同一时刻只有一个中心神经元或一组局部神经元会对当前输入产生积极的响应，并且响应的位置区域倾向于越来越固定，就像通过网络建立了某种与不同输入特征对应的具有意义的坐标系统。所以，与其说神经元在某个位置

收稿时间：2006 – 07 – 06
基金支持：国家自然科学基金(104760061)
作者简介：周巧临(1977 –)，女，助教，主要从事空间数据库、数据挖掘方面的研究.

对输入有积极的响应是对输入输出信号进行了精确的变换，不如说是对输入信号提供的各种模式信息进行了不同的解释。

1.1　问题描述

为了能够得到科学数据时间序列间的相似性聚类[6-8]，将各个时间序列看成高维向量进行分析，采用自组织映射将时间序列映射到不同的聚类中，代表具有不同规律的时间序列模型。所以，利用输出空间中模型向量的距离能够表示输入数据间的相似性，通过分析映射就能够很好地理解输入时间序列的结构。

自组织映射的目的就是使神经元的权系数的形态表示可以间接模仿输入的信号模式。自组织特征映射的学习算法由最优匹配神经元的选择和网络权值系数的自组织两部分组成。

设有输入向量：

$$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$$

对于自组织特征映射网络的输出层神经元 j，则有权系数向量：

$$\boldsymbol{W}_j=(w_{1j},w_{2j},\cdots,W_{nj})^{\mathrm{T}}$$

式中　$j=1,2,\cdots,n$。神经元的输出 y_j 的初始分布可能是随机的，但随着时间的变化，由于输出层神经元有侧向交互的作用，y_j 的分布就会因对环境的自组织而形成“气泡”状，此时神经元权值向量的分布将与各个聚类的中心分布一致。

1.2　距离函数

输入向量与神经元权值向量之间的匹配程度即为输入向量和神经元权值向量之间的相似性。相似性判断距离可以采用明考夫斯基距离、绝对距离(曼哈顿距离)、切比雪夫距离、兰氏距离、夹角余旋、欧几里德距离等方法。

由于科学数据时间序列构成的向量可能达到数千维以上，因此在对时间序列进行相似度计算时，要避免高维向量的累积效应，更加充分地体现出时间序列变化的差异。下面以一个具体的例子来说明高维向量的累积效应。

设定向量 $\boldsymbol{X}_0$、$\boldsymbol{X}_1$、$\boldsymbol{X}_2$ 分别表示三个神经元的权值向量；$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 分别表示科学仿真中位置1和位置2处对应的科学数据时间序列向量，且 $\boldsymbol{X}_0$、$\boldsymbol{A}$、$\boldsymbol{B}$ 的值分别为：

$$\boldsymbol{X}_0=[1.0,\ 1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,0.0]$$
$$\boldsymbol{A}=[1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,1.0,0.9]$$
$$\boldsymbol{B}=[1.1,1.1,1.1,1.1,1.1,1.1,1.1,1.1,1.1,0.0]$$

在采用曼哈顿距离标准时，可判断出 $\boldsymbol{A}$、$\boldsymbol{B}$ 与神经元 $\boldsymbol{X}_0$ 的距离相同。但是，$\boldsymbol{A}$、$\boldsymbol{B}$ 与神经元 $\boldsymbol{X}_0$ 向量的相似性实际上却相差很大。向量 $\boldsymbol{A}$ 与神经元 $\boldsymbol{X}_0$ 向量完全不对称。

在上述例子中采用明考夫斯基距离、绝对距离(曼哈顿距离)、切比雪夫距离、兰氏距离等判断距离的标准，都存在类似的问题。

为了真实地反映高维历史时间序列向量之间的相似度，不仅要考虑向量之间的欧氏距离，还要考虑向量差数列的离散程度，即先将两向量对应的维数相减，得到一个与两向量同维数的差向量，然后再将差向量作为一个数列(该数列称为差数列)，求解差数列的方差，如果获得的差数列的方差大于某一阈值，则说明上述两向量之间的欧氏距离不是由高维向量的累积效应产生的。

考虑到上述判断规则，可以将两个历史时间序列向量的相似度定义为如下形式：

$$\mathrm{Dist}\left\|\boldsymbol{X}_1,\boldsymbol{X}_2\right\|=\frac{\sum_{i=1}^{n}\sqrt{\left(\boldsymbol{X}_{1_i}-\boldsymbol{X}_{2_i}\right)^2}}{n}\times\sum_{i=1}^{n}\left|\sqrt{\left(\boldsymbol{X}_{1_i}-\boldsymbol{X}_{2_i}\right)^2}-\frac{\sum_{i=1}^{n}\sqrt{\left(\boldsymbol{X}_{1_i}-\boldsymbol{X}_{2_i}\right)^2}}{n}\right|$$

式中　$\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 为需要进行比较的两个科学数据时间序列；n 为时间序列包含的时间步；$\boldsymbol{X}_1$ 为科学数据时间序列；$\boldsymbol{X}_2$ 为神经元权值向量。如果定义输入向量为 $\boldsymbol{X}_j$，神经元权值向量为 $\boldsymbol{w}_i$，输入向量与神经元权值向量之间的距离可表示为：

$$d_{ij}=\mathrm{Dist}\left\|\boldsymbol{X}_j,\boldsymbol{w}_i\right\|\qquad i=1,2,\cdots,m$$

2　改进的灰色理论预测模型

灰色系统理论以“部分信息已知，部分信息未知”的“贫信息”不确定性系统为研究对象，主要通过对“部分”已知信息的生成、开发，提取有价值的信息，实现对系统运行规律的正确描述和有效控制。因为科学仿真中存在着时间间隔变化的网格数据，所以得到的时间序列不满足“等间隔”规律，必然会丢失数据的一部分信息。又因为物理现象的仿真结果具有“部分信息已知，部分信息未知”的特点，且系统的数据量庞大，要求处理速度快，所以不适宜采用传统的时间序列分析方法(如Yule-Walker估计、LMS估计(最小二乘法估计)、最大似然估计)进行时间序列的预测。但以上特点充分满足灰色系统理论研究对象的要求，故适宜用灰色理论进行建

模预测。

GM1_1是一种灰色系统模型，它能对时间序列建模，并进行预测。但GM1_1原形中没有考虑对模型参数的优化，使得模型的精确度不高。有很多对模型参数优化的不同方法，其中一种是优化模型中的特解。本文的方法也对模型中的特解进行优化。

GM1_1模型的原形为：

$$\frac{\mathrm{d}\boldsymbol{x}^{(1)}(t)}{\mathrm{d}t}+a\boldsymbol{x}^{(1)}(t)=b \tag{1}$$

式中 $\boldsymbol{x}^{(1)}(t)$ 是原时间序列的 t 步1阶累加算子序列，且：

$$\boldsymbol{x}^{(1)}(t)=\sum_{i=1}^{t}\boldsymbol{x}^{(0)}(i)$$

式中 $\boldsymbol{x}^{(0)}(t)$ 代表t步的原时间序列数据；a 是模型的发展系数；b 是模型的灰色作用量。

一阶累加算子序列的通解为：

$$\boldsymbol{x}^{(1)}(t)=\frac{b}{a}+c\mathrm{e}^{-at} \tag{2}$$

式中 c为GM1_1模型中的特解。

原时间序列(零阶累加算子序列)的通解为：

$$\boldsymbol{x}^{(0)}(t)=c(\mathrm{e}^{-at}-\mathrm{e}^{-a(t-1)}) \tag{3}$$

将 $t=1$ 带入式(3)得到一个特解：

$$\boldsymbol{x}^{(1)}(t)=\left(\boldsymbol{x}^{(1)}(1)-\frac{b}{a}\right)\mathrm{e}^{-a(t-1)}+\frac{b}{a}$$

$$=\left(\boldsymbol{x}^{(0)}(1)-\frac{b}{a}\right)\mathrm{e}^{-a(t-1)}+\frac{b}{a}$$

故有：

$$c=\left(\boldsymbol{x}^{(1)}(1)-\frac{b}{a}\right)\mathrm{e}^{a}$$

将 c 代入式(3)，可得预测模型为：

$$\boldsymbol{x}^{(0)}(t)=\left(1-\mathrm{e}^{a}\right)\left(\boldsymbol{x}^{(0)}(1)-\frac{b}{a}\right)\mathrm{e}^{-at}$$

对GM1_1模型中特解 c 的优化步骤如下：

在得到通解后，暂不计算出特解，而保留 $\boldsymbol{c}$ 未知，进一步计算出 $\boldsymbol{x}^{(0)}(t)$，即把 $\boldsymbol{x}^{(1)}(t)$ 还原得到 $\boldsymbol{x}^{(0)}(t)=c(1-\mathrm{e}^{a})\mathrm{e}^{-at}$，把$t$的所有取值全部列出，得到一个方程组：

$$\begin{cases}\boldsymbol{x}^{(0)}(1)=c(1-\mathrm{e}^{a})\mathrm{e}^{-a}\\ \boldsymbol{x}^{(0)}(2)=c(1-\mathrm{e}^{a})\mathrm{e}^{-2a}\\ \vdots\\ \boldsymbol{x}^{(0)}(n)=c(1-\mathrm{e}^{a})\mathrm{e}^{-na}\end{cases}$$

令 $\boldsymbol{X}=\begin{pmatrix}\boldsymbol{x}^{(1)}(1)\\ \boldsymbol{x}^{(1)}(2)\\ \vdots\\ \boldsymbol{x}^{(1)}(n)\end{pmatrix}$ 和 $\boldsymbol{D}=\begin{pmatrix}\mathrm{e}^{-a}\\ \mathrm{e}^{-2a}\\ \vdots\\ \mathrm{e}^{-na}\end{pmatrix}$，方程组改写成向量的形式得：

$$\boldsymbol{X}=c(1-\mathrm{e}^{a})\boldsymbol{D}$$

用最小二乘法(LMS)估计 c 的取值：

$$\frac{\mathrm{d}(c(1-\mathrm{e}^{a})\boldsymbol{D}-\boldsymbol{X})^{\mathrm{T}}(c(1-\mathrm{e}^{a})\boldsymbol{D}-\boldsymbol{X})}{\mathrm{d}c}=0$$

$$\Rightarrow c=\frac{(\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{X}}{1-\mathrm{e}^{a}}$$

将 c 带入式(4)可得到优化后的预测模型为：

$$\boldsymbol{x}^{(0)}(t)=\frac{(\boldsymbol{D}\boldsymbol{D}^{\mathrm{T}})^{-1}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{X}}{1-\mathrm{e}^{a}}(\mathrm{e}^{-at}-\mathrm{e}^{-a(t-1)})$$

3 试验分析

本文试验采用三维数值模拟数据LaredP进行测试分析。LaredP程序主要用于仿真强激光在高密度等离子体内的传播过程，并通过输出的仿真数据反映其中发生的一系列物理现象。LaredP数据网格规模为80×201×80，在每个时间周期保存一个物理量时需要存储的数据大约为10 M，每次模拟仿真都需要上千时间周期以及诸多物理量。将每个时间周期采样数据称为一个时间步，以电场强度第6时间步时Z轴中心切片数据(201×80)为例，可视化如图1所示。

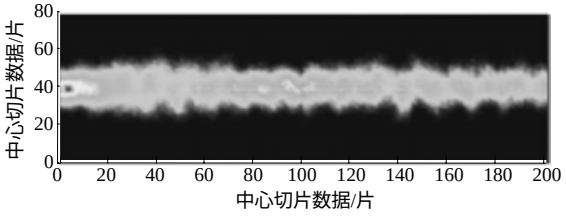

图1 LaredP数据第6时间步Z轴中心切片

在图1所示的图像中，自左向右在等离子体中心轴上选取13个点进行分析。由于仿真过程结果输出的时间步间隔不等，因此，在这13点上电场强度随时间变化产生的时间步序列如图2所示，只有到达标注的时间步时才输出计算结果。

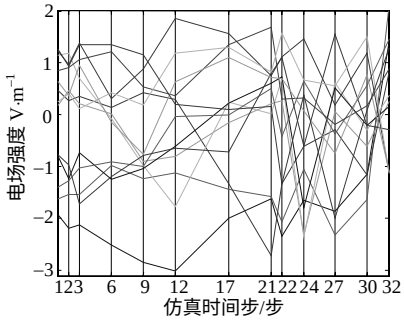

图2 电场强度时间序列集合

由于这13个点的纵坐标都为 40，故以它们的横坐标标记这些点的电场强度时间序列，采用改进的自组织聚类算法对它们进行聚类，得到如图3所示的

聚类结果。

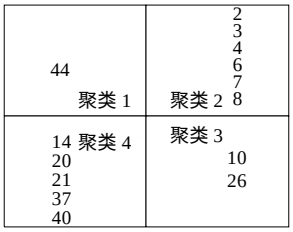

图3　时间序列聚类结果

针对每个聚类选择一个时间序列，采用改进型GM算法进行时间序列预测建模，其建模性能如表1所示。

表**1**　时间序列预测建模性能对比

时间序列/性能	改进型GMLZ		一般型GMBW	
	平均相对误差	平均绝对误差	平均相对误差	平均绝对误差
6(聚类2)	0.158	1.368	0.159	1.371
21(聚类4)	0.106	1.175	3.465	6.814
26(聚类3)	0.179	1.172	0.192	1.175
44(聚类1)	0.167	0.654	0.172	0.704

4　结　论

由试验结果可知，运用改进的距离标准能够充分体现高维时间序列之间的相似性，并克服高维向量的累积误差效应。采用改进型的GM预测方法对LardP数据进行预测建模时，平均相对误差和平均绝对误差都有所改善。根据科学数据的特点，在进行预测前首先对时间序列进行自组织聚类能有效地分离不同类型的时间序列，再针对具有不同规律的科学数据时间序列分别进行预测建模，能极大地提高预测模型的精确度和准确率。

使用灰色理论结合神经网络聚类模型进行科学数据时间序列预测，在时间上和效率上能达到目前实际应用的需求，且具有本身的独特优势，能够分析出数据中不同的变化规律，所以能够取得更好的预测效果。因此，该模型对于科学数据分析具有实际意义和价值。

本文研究工作得到西南科技大学青年预研基金(06ZX3174)资助，在此表示感谢！

参　考　文　献

[1] 陈　虹, 张　侠, 夏　芳, 等. 三维等离子体粒子模拟程序的数据模型和I/O性能改进[J]. 计算机工程与应用, 2004, 40 (9): 104 -107.

[2] Technical Report CS20227, Modeling and querying scientific simulation mesh data[R]. Vermont: Department of Computer Science, University of Vermont, 2002.

[3] YADAV R, N, KALRA P K, JOHN J. Time series prediction with single multiplicative neuron model[J]. Applied Soft Computing, 2007, 7(4): 1157-1163.

[4] LIN Yong-huang , LEE Pin-chan. Novel high-precision grey forecasting model[J]. Automation in Construction, 2007, 16(6): 771-77.

[5] TUCKER A, SWIFT S, LIU X. Variable grouping in multivariate time series via correlation[J]. IEEE Trans Systems, Man, Cybern, Part B: Cybernetics, 2001, 31(2): 235-245.

[6] LIAO T W. Clustering of time series data—a survey, pattern recognition[J]. 2005, 38 (11): 1857-1874.

[7] XIE X L, BENI G. A validity measure for fuzzy clustering[J]. IEEE Trans Pattern Anal Mach Intell, 1991, 13(8): 841-847.

[8] LIAO T W, TING C F, CHANG P C. An adaptive genetic clustering method for exploratory mining of feature vector and time series data[J]. Int J Production Res, 2006, 44(14): 2731-2748.

[9] HAN J, KAMBER M. Data mining: Concepts and techniques[R]. San Francisco: Morgan Kaufmann, 2001.

[10] ROMONI M, SEBASTIANI P, COHEN P. Bayesian clustering by dynamics[J]. Mach Learn, 2002, 47(1): 91-121.

编　辑　熊思亮