

两种样本输入方式下基于 GRNN 的 日长变化预报结果的比较

张晓红, 王琪洁, 朱建军, 张昊

(中南大学地球科学与信息物理学院, 长沙 410083)

摘要: 针对广义回归神经网络用于日长变化预报过程中, 样本的输入方式对预报结果的影响进行了研究。采用 2 种输入方式: 即样本按不同跨度输入以及按连续输入, 对日长变化进行预报。最终证明不同的样本输入方式对日长变化预报精度的影响较大, 样本按跨度输入在超短期预报中预报精度较高, 样本采用连续输入的方式在短期和中期预报中预报精度较高。

关键词: 广义回归神经网络; 日长变化预报; 输入方式; 相关性

中图分类号: TP274, P183.3⁺¹

1 引言

高精度的地球自转参数预报具有重要的科学意义和实际应用价值, 其中表征地球自转速率变化的参数是日长变化。至今已检测到的日长变化的时间尺度从数十年到几小时不等^[1]; 日长的 10 年尺度起伏是由地球内部的核幔耦合效应引起的, 日长的年际变化与厄尔尼诺和南方涛动事件 (ENSO) 以及大气角动量 (AAM) 变化有着密切的关系, 日长的季节性变化主要源于大气环流的季节性变化, 日长的亚季节性变化主要是由大气的高频振荡和潮汐波动引起的^[2-5]。

日长变化包含了复杂的非线性因素, 尤其是受到大气、海洋变化影响时, 其预报难度加大, 因此采用非线性神经网络预报的方法更为合理。最先将神经网络技术用于地球自转参数预报的是 Egger, 他的研究验证了神经网络用于地球自转参数 (ERP) 预报的有效性, 而且显示了神经网络技术在解决非线性问题上蕴藏的巨大潜力^[6-7]。

2002 年, Schuh 用斯图加特神经网络模拟软件 (即 BP 神经网络) 对地球自转参数进行了中长期预报, 验证了神经网络是预报地球自转参数较适合的工具, 同时他也指出, 神经网络的训练过程要在合适的时间停止^[8]。2005 年, Kalarus 和 Schuh 又对地球自转参数做了进一步研究, 他们采用最小二乘和神经网络结合的方法进行预报, 最小二乘用于线性序列的预报, 神经网络或自回归模型用于非线性残差序列的预报, 最后把 2 种模型预报的结果相加, 得到了比单独使用 1 种方法预报更好的效果^[9]。2007 年, 王琪洁等用 BP 神经网络对地球自转参数做了短期和中长期预报, 并将大气角动量引入神经网络预测模型, 对神经网络拓扑结构的正确选择也做了一定的研究; 她们还应用 BP 神经网络对地球自转变化的实时快速预报进行

收稿日期: 2011-01-17; 修回日期: 2011-02-21

资助项目: 国家自然科学基金委员会与中国科学院天文联合基金 (No.10878026)

了研究, 并获得成功^[10-11]。

但是 BP 神经网络存在以下缺点: 易陷入局部极小值, 网络拓扑结构难以确定, 泛化能力不高等^[12-13]。随着神经网络技术的发展, 新的神经网络模型逐渐建立, 广义回归神经网络 (GRNN) 是继 BP 神经网络之后又一广泛应用于拟合、预测、模式识别等领域的神经网络模型, 它由 Specht 于 1991 年提出。GRNN 是基于统计理论建立的模型, 它只有 3 层, 不需要每次对初始数据循环训练, 只要计算 1 次便可找到最优解, 即使在样本数据很少的情况下, GRNN 也能达到较好的效果, 从而大大提高工作效率, 这是 GRNN 最重要的优势, 也是 Specht 建立 GRNN 模型的初衷。他指出, GRNN 特别适用于解决实时性要求较高的问题^[14], 这也是我们采用 GRNN 用于日长变化预报的主要原因。

神经网络用于日长变化预报的过程中, 训练样本的输入方式通常有 2 种: 一是按跨度输入, 即按间隔 $i d (i = 1, 2, 3, \dots)$ 的方式输入, 二是按连续方式输入。按跨度输入的优点是对数据进行抽样处理, 减少了计算量, 在数据量庞大的情况下, 可以快速实现预报。但由于跨度较大时, 间隔较远的日长值之间相关性很小, 而且容易忽略掉中间短周期振荡对日长变化的影响, 这样神经网络很可能忽略一些重要的先验信息。为了进一步研究这 2 种输入方式的预报效果, 本文利用 20 a 的日长变化数据, 分别采 2 种输入方式进行预测, 给出了各自的预报结果并进行了对比。

2 广义回归神经网络的基本原理

2.1 广义回归神经网络的理论基础

GRNN 的理论基础是非线性回归分析。设 $f(x, y)$ 为非独立变量 y 和独立变量 x 的联合概率密度, 要计算 y 相对于 x 的回归分析实际上是计算具有最大概率值的 y 。设 X 为变量 x 的观测值, 则 y 相对于 X 的条件均值为:

$$E(y|X) = \hat{y}(X) = \frac{\int_{-\infty}^{+\infty} y f(x, y) dy}{\int_{-\infty}^{+\infty} f(x, y) dy}, \quad (1)$$

其中, $f(x, y)$ 可以由训练样本估计得到, 其计算公式为:

$$f(x, y) = \frac{1}{n(2\pi)^{(p+1)/2}\sigma^{p+1}} \sum_{i=1}^n \exp\left[-\frac{(X - X_i)^T(X - X_i)}{2\sigma^2}\right] \exp\left[-\frac{(y - Y_i)^2}{2\sigma^2}\right], \quad (2)$$

其中 n 为样本数, p 为随机变量 x 的维数, X_i 和 Y_i 分别为随机变量 x 、 y 的样本观测值, σ 为平滑参数, 即核函数 (激励函数) 的宽度。

将 (2) 式代入 (1) 式得:

$$\hat{y}(X) = \frac{\sum_{i=1}^n \exp\left[-\frac{(X - X_i)^T(X - X_i)}{2\sigma^2}\right] \int_{-\infty}^{+\infty} y \exp\left[-\frac{(y - Y_i)^2}{2\sigma^2}\right] dy}{\sum_{i=1}^n \exp\left[-\frac{(X - X_i)^T(X - X_i)}{2\sigma^2}\right] \int_{-\infty}^{+\infty} \exp\left[-\frac{(y - Y_i)^2}{2\sigma^2}\right] dy}, \quad (3)$$

再根据 $\int_{-\infty}^{+\infty} ze^{-z^2} dz$ 变换为:

$$\hat{y}(X) = \frac{\sum_{i=1}^n y_i \exp \left[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2} \right]}{\sum_{i=1}^n \exp \left[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2} \right]} \quad (4)$$

条件均值 $\hat{y}(X)$ 为所有训练样本的加权平均值, 观测值 Y_i 的权重因子为相应样本 X_i 与 X 的 Euclid 距离平方的指数。

2.2 广义回归神经网络的结构

GRNN 是一种径向基函数 (RBF) 神经网络, 由 3 层组成, 即输入层、隐含层 (分为模式层和求和层) 和输出层。它与 RBF 神经网络的区别在于输出层的数据处理方式不同。GRNN 的结构如图 1 所示。

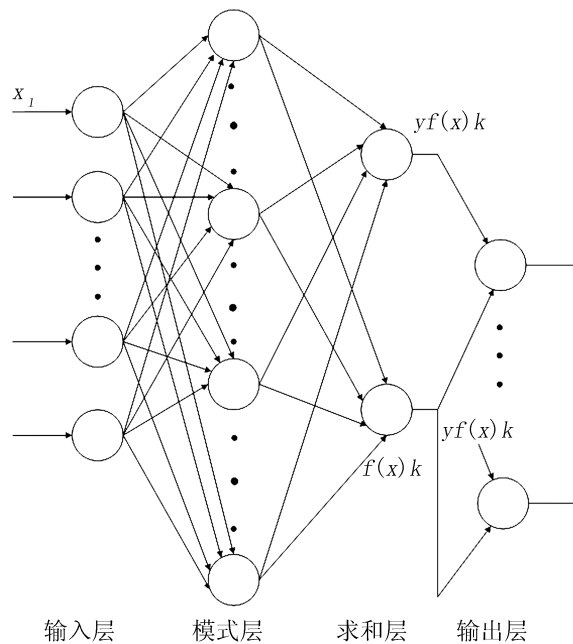


图 1 GRNN 神经网络结构

第 1 层为输入层, 神经元数目等于输入向量的维数。

网络的第 2 层为模式层和求和层。模式层中神经元个数等于训练样本数, 不同的神经元对应不同的样本。设样本中心为 X_i , 神经元的传递函数为:

$$p_i = \exp \left[-\frac{(X - X_i)^T (X - X_i)}{2\sigma^2} \right] \quad (5)$$

其中 σ 为光滑因子, σ 越大则激励函数越平滑。求和层中包括 2 种类型的神经元^[15], 一种计算 (4) 式中的分子部分, 即图 1 中对应的 $yf(x)k = \sum_{i=1}^n yp_i$, 其中 k 是一个常数, 另一种神经

元计算 (4) 式中的分母部分, 即图 1 种中对应的 $f(x)k = \sum_{i=1}^n p_i$ 。

第 3 层 (输出层) 的最终输出值简写为:

$$\hat{y}(X) = \frac{\sum_{i=1}^n y p_i}{\sum_{i=1}^n p_i} \quad (6)$$

3 广义回归神经网络用于日长预报

3.1 数据预处理

文中所用日长资料取自国际地球自转和参考服务组织 (IERS) 的 EOP C04 序列, 时间跨度为 1980—2010 年, 每日 1 个值。首先对日长序列中周期为 5 d~18.6 a 的 62 个固体地球带谐潮项进行扣除, 日长变化序列中的周期项 (包括周年项, 1、2 次谐波, 周期长度等于 1、1/2 或 1/3 数据序列长度的项) 通过最小二乘拟合的方法确定, 剩下的残差序列文中用 LODR¹ 表示, 这一部分的预报我们采用广义回归神经网络的方法, 最后将线性预报结果和非线性预报结果相加即为最终的日长值。在进行日长变化预报前, 先对日长变化残差序列进行归一化处理, 使其在 [-1,1] 之间。

3.2 建模和预报

日长变化的年际、季节、亚季节变化与大气、海洋的质量重分布及一些重要的自然现象有着密切的关系。以跨度为 300 d 的日长变化预报为例, 输入样本的取样间隔为 300 d, 那么这 300 d 内的大气非线性变化信息并没有参与网络的训练, 采用这样的样本作为训练数据, 神经网络记下的可能只是较为微弱的相关性, 并不能获得相近数据之间的全面信息, 这势必影响新样本的泛化能力。

本文选取 1980 年 1 月 1 日至 1999 年 12 月 31 日的日长残差序列进行神经网络的训练, 预报 2000 年至 2001 年的 365 个日长残差值。首先将训练样本按 80%、10%、10% 的比例分为训练段、验证段和测试段。为了研究不同样本输入方式对预报结果的影响, 按以下 2 种方式输入样本:

(1) 按跨度 i 输入

即相应于时间历元 t 、跨度为 i ($i = 1, 2, 3, \dots$) 的预报, LODR 数据的输入方式为: LODR($t - s \times i$), \dots , LODR($t - 3 \times i$), LODR($t - 2 \times i$), LODR($t - 1 \times i$), 式中的 s 为网络输入层的神经元个数。

(2) 连续输入

即相应于时间历元 t 、跨度为 i ($i = 1, 2, 3, \dots$) 的预报, LODR 数据的输入方式为: LODR($t - i - s + j$), $j = 1, 2, 3, \dots, s$, 式中的 s 同样为网络输入层的神经元个数。神经元个数等于训练样本数, 不同的神经元对应不同的样本, 这也是 GRNN 区别于 BP 算法之处。

¹这里与通常所说的 LODR 有所不同, 通常 LODR 表示经过固体地球带谐潮修正后的日长变化序列

隐含层节点的个数从 2 取到 10, 每个节点计算 5 次, 取均方根误差 (RMSE) 平均值最小时的神经网络模型作为最终的预报模型。设网络输出为 t_i , 目标输出为 e_i , 则均方根误差的数学表达式为:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (t_i - e_i)^2} \quad (7)$$

GRNN 每次计算需要调节的是光滑因子, 用 σ 表示。首先根据式 $\sigma_{i+1} = \sigma_i + \Delta\sigma$, 设置较大的 $\Delta\sigma$ 值, 如 0.05, 在较大的范围内进行计算, 找到一个使 RMSE 取较小值的光滑因子的范围; 将范围缩小后, 将 $\Delta\sigma$ 设置成较小的值, 如 0.01, 在新的较小的范围内调节光滑因子, 最终找到使 RMSE 最小的光滑因子。

4 实验结果

根据上述建模方式, 采用 2 种不同的输入方式对日长变化进行 1, 2, 3, ..., 10, 15, 20, 25, 30, 60, 90, ..., 390 d 的预报, 将计算结果列于表 1 中。表 1 中第 4 列“精度改善百分比”表示按连续方式输入预报的精度相对于按跨度为 i 的方式输入预报的精度改善情况, 其中负号表示没有改善。

为了使结果更加直观, 我们给出了比较图, 见图 2。图例中 method 1 表示按跨度 i 的方式输入, method 2 表示按连续的方式输入。

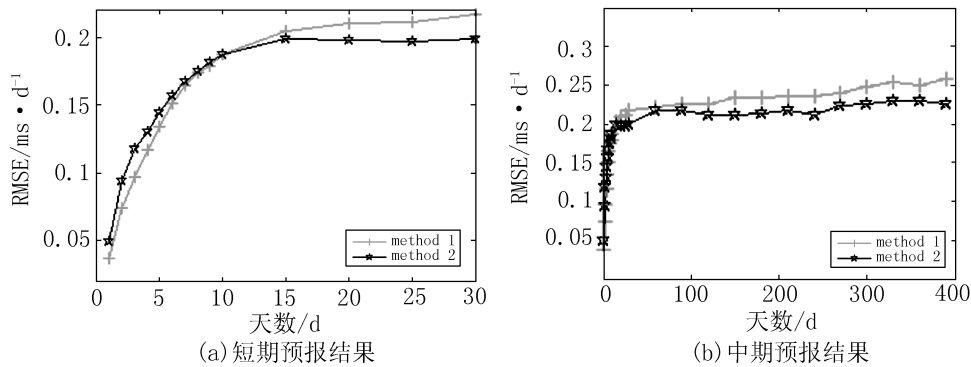


图 2 不同输入方式日长预报精度比较

通过表 1 和图 2, 可以发现: 对于 1~10 d 的预报, 按连续方式输入的预报精度低于按跨度 i 输入的预报精度。从第 15 天开始, 按连续输入方式的预报精度均优于按跨度 i 输入的预报精度, 并逐渐拉开距离, 精度改善最大值为 12%。这也证明了我们实验前的推测, 如果输入方式选择按跨度的方式, 当跨度较大时, 间隔较远的日长值之间相关性很小, 神经网络预报结果的准确性会降低。

5 总结与展望

本文基于神经网络能够主动向环境学习的特性, 提出 GRNN 应采用连续输入的方式进行日长变化预报, 这种输入方式有利于神经网络记下足够多的日长变化信息。它能更好地利

表 1 不同输入方式日长预报精度的比较

天数/d	GRNN (跨度为 i 输入) /ms·d ⁻¹	GRNN (连续输入) /ms·d ⁻¹	精度改善百分比
1	0.037	0.049	-24%
2	0.074	0.094	-27%
3	0.097	0.118	-22%
4	0.117	0.130	-11%
5	0.134	0.144	-7%
6	0.151	0.157	-4%
7	0.164	0.167	-2%
8	0.174	0.175	-1%
9	0.179	0.181	-1%
10	0.187	0.187	0%
15	0.204	0.199	+2%
20	0.210	0.198	+6%
25	0.211	0.197	+7%
30	0.217	0.199	+8%
60	0.222	0.217	+2%
90	0.226	0.217	+4%
120	0.226	0.211	+7%
150	0.233	0.212	+9%
180	0.234	0.213	+9%
210	0.236	0.217	+8%
240	0.236	0.212	+10%
270	0.240	0.223	+7%
300	0.247	0.226	+9%
330	0.254	0.229	+10%
360	0.250	0.229	+9%
390	0.257	0.226	+12%

用大气的季节变化、潮汐波动及大气的高频振荡等先验信息。GRNN 本质上是基于统计理论的,这就更需要充分利用观测值的先验信息。本文使用的实验数据是 1980 年至 2001 年的日长变化资料,在这期间,发生了 1982~1983 年、1997~1998 年 2 次厄尔尼诺事件以及 1998~1999 年的拉尼娜事件,这些事件都与日长的年际变化密切相关,采用连续输入的方式能更充分地研究日长变化多种时间尺度上的变化信息。实验证明,这种预报方法在跨度较长的日长预报中有很好的预报精度,在跨度大于 20 d 的日长变化预报中,精度改进在 10% 左右。

实验证明:这种输入方式对本实验采用的训练样本有效,但是对其它实验数据是否有效还有待进一步研究。在接下来的研究中可以采用新的日长序列进行日长变化预报,也可以对更长跨度的日长预报进行实验分析,以进一步探讨本文结论。我们也可以采用 2 种输入方式的结合方法用于日长变化预报,即在超短期 1~10 d 的日长变化预报中采用按跨度的方式输入,在大于 10 d 的短期和中期预报中采用本文提出的方法。

参考文献:

- [1] Lambeck K. The Earth's Variable Rotation. Cambridge: Cambridge University Press, 1980
- [2] 周永宏, 郑大伟, 廖新浩. 测绘学报, 2001, 30(4): 288
- [3] 闫昊明, 钟敏, 朱耀仲. 天文学报, 2005, (46): 452
- [4] 顾晖, 郑大伟. 中国科学院上海天文台年刊, 1992, (13): 13
- [5] 杨志根. 中国科学院上海天文台年刊, 1992, (13): 36
- [6] Egger D. Allgemeine Vermessungsnachrichten(AVN), 1992, 11/12, S: 517
- [7] Egger D, FRÖHLICH H. Allgemeine Vermessungsnachrichten(AVN), 1993, 10, S: 366
- [8] Schuh H, Ulrich M, Egger D *et al.* Journal of Geodesy, 2002, 76: 247
- [9] Kalarus M, Schuh H, Kosek W. Geophysical Research Abstracts, 2005.7:01753.
- [10] 王琪洁, 廖德春, 周永宏, 等. 天文学报, 2008, (49): 93
- [11] 王琪洁. 博士论文. 上海: 中国科学院上海天文台, 2007
- [12] 焦李成. 神经网络系统理论. 西安: 西安电子科技大学出版社, 1996: 17
- [13] 阎平凡, 张长水. 神经网络与模拟进化计算. 北京: 清华大学出版社, 2005: 19
- [14] Specht D F. IEEE Transactions on Neural Networks, 1991, (2): 568
- [15] Ben-Nakhi A E, Mahmoud M A. Energy Conversion and Management, 2004, (45): 2127

Comparison of Results of Forecasting LOD Based on GRNN with two Input Architectures

ZHANG Xiao-hong, WANG Qi-jie, ZHU Jian-jun, ZHANG Hao

(School of Geosciences and Info-Physics, Central South University, Changsha 410083)

Abstract: A comparison of the result of two input architectures, sample in different lead time and sample in sequence, in forecasting LOD by General Regression Neural Network (GRNN) model is presented. Results show that two different input architectures have obvious different influence on accuracy of the LOD prediction. The former leads to higher accuracy in ultra short-term prediction, while the latter leads to higher accuracy in short-term and medium-term predictions.

Key words: General Regression Neural Network(GRNN); prediction of LOD; input architecture; relativity