基于数据预处理灰色神经网络组合和集成预测

严修红[1,2],许伦辉[2],董世畅[1]
(1. 顺德区容山中学,广东 顺德 528303;2. 江西理工大学 机电工程学院,江西 赣州 341000)

摘 要:当研究的系统扰动因素过大或系统行为在某个时间点发生突变,出现严重扰动系统的异常数据时,提出不应直接按原始数据建模预测,而应根据实际情况适当地对数据预处理. 提出了基于数据修正的改进型灰色神经网络组合和集成预测,并根据南昌火车站旅客发送量时间序列建立了多个模型,从模型预测效果对比中说明数据修正+改进型灰色模型和改进型灰色神经网络+灰色神经网络组合和集成确实能提高预测精度. 另外,修正数据要把握一个度,不能修正全部数据,只能修正较异常的数据,要在数据的趋势性和预测的灵敏性间取得平衡.

关键词:时间序列预测;灰色神经网络;组合预测
中图分类号:U491.14 **文献标识码**:A **文章编号**:1673-4785(2007)04-0058-05

Grey neural network and integrated forecasting based on preprocessed data

YAN Xiu-hong[1,2], XU Lun-hui[2], DONG Shi-chang[1]
(1. Rongshan Middle School of Shunde County, Shunde 528303, China; 2. Institute of Electromechanical Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)

Abstract: When a system disturbance is too great or a sudden change occurs, the resulting abnormal data can severely disturb the forecasting system. In this situation, running a forecasting model before abnormalities in the original data are identified produces misleading results. In this paper, an improved grey neural network forecasting model and integrated forecasting method are proposed on the basis of data modification. Several forecasting models were tested based on time sequences of passenger volume in Nanchang Railway Station. After comparing model predictions with real data, it became clear that prediction accuracy is considerably improved with revised data, or an improved grey model, or a combined grey neural network. But the data modification must be done properly. Not all data should be modified, it is only necessary to modify abnormal data in order to maintain balance between the data tendency and forecasting sensitivity.

Keywords: time series forecasting; grey neural network; combined forecasting

现实生活中有很多时间序列,由于条件的限制,获得的数据不会很多,而且数据间的关系往往很复杂,很多不是线性关系,如用传统线性预测方法或基于线性变换的方法预测,一般来说不能达到很好的效果,若用基于数理统计的回归方法去预测,由于这类方法需要大量数据,故往往因数据不足而精度不高,可靠性不强. 由于灰色预测具有少量数据建模和累加生成可增加历史数据的规律性的特点,而神经网络具有良好的逼近任意非线性函数的优势,因此这两种方法在预测领域特别是预测上述这类时间序列效果较好,且灰色预测模型与神经网络预测互为取长补短,具有优势互补性,因此二者组合进行预测可以达到提高预测精度的目的. 同时根据组合预测理论,将它们组合起来,还可增加预测结果的可靠性和稳定性,减少单个预测的风险性;因此结合灰色系统思想与神经网络构成灰色神经网络成为近几年一个研究热点,但灰色神经网络算法大部分灰色模型

收稿日期:2006-09-30.
基金项目:国家自然科学基金资助项目(60664001);江西省自然科学基金资助项目(0511030).

采用传统的 GM (1,1) 与神经网络相组合,由于传统的 GM (1,1) 模型固有的缺陷性,导致预测精度不高,且大部分算法本质上都是建立原始数据的拟合模型,最大限度地提高拟合精度,认为模型的拟合精度越高,预测效果越好. 实际上,模型的拟合精度不等于预测精度. 拟合精度高并不意味着预测精度也一定高. 因此,即使找到了高度拟合曲线,也不代表未来的预测值一定可靠. 由于事物的复杂多样性,系统往往受各种偶然因素的影响而表现得错综复杂,扰动因素非常大,且获得数据往往受条件所限,并不十分可靠,信号数据中可能含有太多的虚假信息和噪声干扰,另外过分地追求拟合将使模型更加复杂,导致预测模型过于适应数据以致于适应噪声,造成预测结果不准确,反而降低了模型的适应能力和推广能力,可见,模型的拟合精度固然重要,但过分追求拟合精度并将其作为预测效果的评价标准是欠妥当的,模型的预测能力不仅是对历史数据的拟合能力,更重要的是模型的适应能力和预测推广能力,如何提高模型的预测能力和预测精度是值得深入探讨的问题。

1　数据修正和预处理

1.1　数据修正和预处理提出的背景

为提高模型的预测能力和预测精度,增加预测的稳定性和预测结果的可靠性,在以上分析的基础上,提出当所研究系统扰动因素过大或系统行为在某个时间点发生突变,当系统行为数据序列存在严重扰动系统的异常数据时,此时系统原始数据不能正确反映系统的真实变化规律,若继续按原始数据建模,不管是用什么样的模型,预测精度都不高,由于系统本身受到某种冲击和干扰而失真,如果努力地去拟合原始数据中的每一数据(包括异常数据),会对系统的发展趋势作出错误的估计,反而会降低预测结果的准确性和可靠性,故不特别着重对历史数据的拟合,而是去修正失真的异常数据,设法排除系统行为数据受到的冲击和干扰,还数据以本来面目,从而提高预测的精度. 异常数据的修正方法有历史数据平稳化、差分处理等,这里采用 RBF 网络对数据进行修正和预处理.

1.2　基于 RBF 网络的数据修正和预处理方法

1.2.1　失真数据的查找

上述数据预处理方法有时效果欠佳,本文采用神经网络查找和修正失真数据,填补空缺数据[1]. 所谓失真数据是指由于某些偶然的因素或特定的情况,造成了某一年或某几年的数据出现了大的转折(如 2003 年突如其来的"非典"造成火车站旅客运输量锐减),这样的数据对未来几年预测的参考价值不大,反而会降低预测结果的准确性和可信性,称这样的数据为失真数据,因此必须查找出失真数据并根据实际情况对其进行修正. 查找方法和步骤为

1) 检验历史各数据是否比前后年份的数据都大(或小),从而判断出该数据是否处于波峰(或波谷),从而把处于波峰或波谷的数据认为是第 1 种类型的失真数据.

2) 建立一个单输入单输出的 RBF 网络,它的输入输出分别为年份和相应年份的数据.

3) 利用归一化后的历史数据(将上述得到的第 1 种类型的失真数据排除在外)对 RBF 网络进行训练. 其中归一化的方法为:对每一个样本输入向量 P 和目标值 t,采用公式

$$p(k) = \frac{p(k) - p_{\min}}{p_{\max} - p_{\min}}, t(k) = \frac{t(k) - t_{\min}}{t_{\max} - t_{\min}} \quad (1)$$

把它归一化为(0,1)内的数.

4) 将归一化后的历史数据中各年份作为输入变量输入到已经完成训练的 RBF 网络,这样就可以得到一组对应的输出值,然后再把这组输出值分别与相应各年份的实际值进行比较,误差超过 5 %的即可认为该年份的数据为失真数据,这是第 2 种类型的失真数据.

1.2.2　失真数据的修正

将查出的所有失真数据年份归一化后作为输入变量输入到已经完成训练的 RBF 网络,这样得到的网络输出反归一化后就可以作为该年份数据修正值. 把该年份数据修正值替代原始数据中的对应的异常数据,其他原始数据不变,则得到修正后的实际值. 这样修正后的实际值在一定程度上排除了系统受到的冲击和干扰,还原了系统的本来面目.

2　多个改进型灰色神经网络的组合和集成模型算法

对数据修正后,可利用修正后的实际值建立多个改进型灰色神经网络,得到多个预测值,最后由这多个预测值进行组合,得出最佳预测值,建立多个改进型灰色神经网络的组合和集成模型算法.

算法步骤为

1) 输入原始数据序列;

2) 用 RBF 网络查找和修正原始数据序列的异常数据,排除扰动因素和噪声干扰;

3) 建立改进型 GM(1,1)模型,再根据模型预测得到改进型 GM(1,1)预测值,用改进型 GM(1,1)

提取趋势性因素,把预测值作为第4步RBF网络的输入;

4)建立改进型灰色神经网络,即用RBF神经网络寻求改进型GM(1,1)预测值与修正后实际值的映射关系,并得出预测值;

5)按上述方法建立另外的改进型灰色神经网络,得到另一组预测值;

6)多组灰色神经网络的预测值再用BP网络进行组合和集成,得到最终预测值.

3 应用实例及模型对比分析

南昌铁路车站旅客发送量见表1实际值[2],现以1996～2003年的实际人数为历史数据按上述算法预测2004年的数据,并对结果进行分析.

表1 实际值与各个模型预测值对照表

Table 1 The contract table between the real values and the forecasting values of models 万人

年份	实际值	修正后的实际值	模型1预测值	模型2预测值	模型3预测值	模型4预测值
1996	508	552.3	552.3	552.3	552.3	549.2
1997	666	630.6	680.2	673.4	630.6	637.8
1998	747	747.0	738.7	731.3	739.1	744.2
1999	807	807.0	802.2	794.1	804.2	806.4
2000	795	889.4	871.2	862.4	875.4	878.1
2001	943	943.0	946.1	936.5	953.2	961.7
2002	1 013	1 013	1 027.4	1 017.0	995.5	1 007.5
2003	968	1 089.6	1 115.8	1 104.4	1 076.7	1 086.9
2004	1 148	1 148.0	1 211.7	1 199.3	1 163.7	1 141.3

3.1 失真数据的修正

建立单输入单输出RBF网络查找和修正失真数据,它的输入输出分别为归一化后的年份和相应的归一化后的年份数据,观察表1的实际人数可知,2000年和2003年的人数处于波谷,属于第1类失真数据,把其他数据归一化后作为训练数据,把所有年份输入后得到一组对应的输出值,然后再把这组输出值分别与相应各年份的实际值进行比较,误差超过5%的有1996、1997、2000、2003年的数据,相对误差分别为-8.71%、5.31%、-11.87%、-12.57%,从而根据拟合值分别修正为552.3、630.6、889.4、1 089.6,从而得出修正后的实际值,实际值与预测值对照见表1.

3.2 建立第一个灰色神经网络(GANN)模型及结果分析

3.2.1 建 模

为便于对比分析预测效果,本文建立多个模型:

模型1 基于原始数据的传统GM(1,1)预测[3].

数据处理前,若直接按原始数据建立传统GM(1,1)进行预测,先用1996～2003年的数据建立8维GM(1,1)模型,模型为

$$\hat{x}^{(0)}_{(1)} = 508,\ \hat{x}^{(0)}_{(k)} = 10\ 251(1 - e^{-0.0653})e^{0.0653(k-1)},\ k = 2,3,\cdots: \quad (2)$$

根据表1拟合值和预测值,得到预测相对误差为-4.86%.

模型2 基于数据修正的传统GM(1,1)预测.

对修正后的实际值建GM(1,1)模型,得到拟合值和预测值,预测相对误差为4.4%,可看出经数据修正后,预测精度有所提高.

模型3 基于数据修正的灰色马尔可夫预测.

若在模型2的拟合值和预测值的基础上再作马尔可夫残差修正[4],即先建立8维GM(1,1)模型预测后,再对残差序列进行马尔可夫预测,模型为

$$\hat{x}^{(0)}_{(1)} = 508,$$

$$\hat{x}^{(0)}_{(k)} = 10\ 251(1 - e^{-0.0653})e^{0.0653(k-1)} + \mathrm{sgn}(i) \cdot 134.4624(1 - e^{-0.1659})e^{0.1659(k-1)},\ k = 2,3,\cdots:$$

式中:符号函数

$$\mathrm{sgn}(k) = \begin{cases} 1, x^{(0)}(k) - \hat{x}^{(0)}(k) > 0, \\ 0, x^{(0)}(k) - \hat{x}^{(0)}(k) = 0, \\ -1, x^{(0)}(k) - \hat{x}^{(0)}(k) < 0, \end{cases}$$

式中:$\mathrm{sgn}(9) = 1$. (3)

根据拟合值和预测值,预测相对误差为-1.36%,在模型2的基础上精度又有所提高.

模型4 基于数据修正的灰色马尔可夫-RBF网络预测

在模型3的基础上建立模型4,即把修正数据后的灰色马尔可夫模型的拟合和预测结果归一化后作为RBF网络的输入,修正后的实际值作为目标输出,采用Matlab的函数solverb建立RBF网络,其中训练目标误差定为0.002,径向基分布常数取

0.7,仿真得到的拟合值和预测值见表 1,预测值反归一化后为 1 141.3,相对误差为 0.59 %;精度已经非常高,预测效果非常好.

3.2.2　模型预测结果对比分析

模型 1～4 的预测误差见表 2,对比模型 1、2 可知,作了数据修正的 GM(1,1)模型比没进行修正的预测效果好,说明修正了异常数据确实能提高精度;对比模型 2、3 可知,说明改进型 GM(1,1)模型比传统 GM(1,1)精度更高;对比模型 3、4 可知,说明改进型灰色神经网络比改进型灰色模型精度更高;预测效果已经很好了,作者也曾在模型 3 的基础上按同样方法去建同样结构的 RBF 网络,但以原始数据为输出样本目标值,以模型 3 的拟合值作为输入样本训练 RBF 网络,若训练目标误差定得较小,如定 0.002,径向基分布常数取 0.7,仿真得到的拟合值与实际值一样,但 2004 年的预测值为28 733,与实际值 1 148 差得太远,根本没有预测价值;这是由于过分拟合历史数据导致过分适应系统数据中包含的噪声信息,而最后一个数据恰巧又是个特别异常的数据(2003 年的罕见的"非典"导致火车站旅客量锐减);如把训练目标误差定大一些(如定为 0.5),则精度更高一些,预测值为 955.29,相对误差为 16.79 %,与修正后的实际值为输出目标的精度相差甚远,可见对异常数据修正是非常必要的.

表 2　4 个模型预测效果对比表

Table 2　The contract table of four models forecasting results

模型	预测误差/ %
模型 1	- 4.86
模型 2	- 4.47
模型 3	- 1.36
模型 4	0.59

3.3　建另一个灰色神经网络(GANN)

先对数据滑动平均处理[5],再选最佳定解条件[6],得出一组拟合值和预测值(见表 3 滑动定解 GM 预测值).建立单输入单输出的 RBF 网络,把改正型滑动定解模型的拟合值和预测值归一化后的值作为 RBF 网络的输入,RBF 网络输出的值反归一化后即为最终拟合值和预测值,训练样本的输入为 1996～2003 年的改正型滑动定解 GM 的拟合归一化后的值,对应的目标值为用 RBF 网络修正后的实际人数值(见表 3),同样采用 Matlab 的函数 solverb 建立 RBF 网络,其中训练目标误差定为 0.02,径向基分布常数取 1,仿真后得到的 RBF 网络中的 radbas 层的神经元个数为 2,1996～2004 年的预测值为1 140.3,与实际值的相对误差为 0.67 %,与实际值非常接近,预测效果相当好.另外,作者还曾尝试用滑动平均处理后的值作为 BP 网络输出目标,则预测误差较大,原因可能是每个原始数据都进行了修正,反而失去了系统的动态信息,可见,数据修正只能对严重扰动数据进行修正,对其他原始数据不能修改,否则会掩盖系统的自身动态发展规律.

3.4　神经网络的组合和集成

再把第 1 个灰色神经网络的预测值与第 2 个灰色神经网络的拟合值和预测值作为神经网络的输入,用 RBF 网络修正后的实际值作为输出,建立结构为 2 - 25 - 1 的 BP 网络,其中隐含层传输函数用 tansig 函数,输出层传输函数用 purelin 函数,权值修正采用 Matlab 中的含动量规则的 BP 学习规则函数 learnbpm,自适应学习速率取 0.05,动量因子取 0.95,初始权值和初始偏值随机取(0,1)的值,初始权值和偏值修正矩阵取零矩阵,检验函数为网络修正权值后误差平方和,训练目标小于 0.000 1,要先对数据归一化,对每一个样本输入向量 P 和目标值 t,采用 $p(k)=\frac{p(k)-p_{\min}}{p_{\max}-p_{\min}}$,$t=\frac{t-p_{\min}}{p_{\max}-p_{\min}}$ 的方法对它归一化,最后再反归一化还原数据,利用 Matlab7.0 编程仿真多次,由于初始权值和初始偏值随机取值,所以检验样本 2004 年的预测数据每次运行结果稍有不同,但与 2004 年的实际值的相对误差都不超过 0.1 %,本文取了相对误差较小的一次结果为:经过 10 000 步训练,从 1996～2004 年的拟合值与预测值结果见表 3,其中 2004 年的预测值为 1 147.6,与实际值相对误差为 0.01 %,可见拟合值非常接近实际值,神经网络集成预测能大大提高预测精度,预测非常精确,预测能力非常强.

表 3　南昌车站旅客发送量及其预测值、相对误差

Table 3　The forecasting value and realtive error of passenger transmission volume in Nanchang railway station　　万人

项目	数据采集的年份								
	1996	1997	1998	1999	2000	2001	2002	2003	2004
实际人数	508.0	666.0	747.0	807.0	795.0	943.0	1 013.0	968.0	1 148.0
修正后的实际值	552.3	630.6	747.0	807.0	889.4	943.0	1 013.0	1 089.6	1 148.0
滑动定解 GM 模型预测值	547.5	675.9	727.1	782.2	841.4	905.2	973.7	1 047.5	1 126.8
第 1 个 GANN 预测值	549.2	637.8	744.2	806.4	878.1	961.7	1 007.5	1 086.9	1 141.3
第 2 个 GANN 预测值	540.1	671.3	731.6	799.1	872.0	947.7	1 021.7	1 088.3	1 140.3
本文最终预测值	552.2	632.8	749.0	804.8	887.5	946.5	1 009.1	1 089.6	1 147.6

3.5 模型对比及结果分析

各个模型的误差对比见表4.

表4 模型误差对比表

Table 4 The contract table of models error %

模型	相对误差
传统 GM(1,1)	4.86
第1个灰色神经网络预测	0.59
第2个灰色神经网络预测	0.67
灰色神经网络集成	0.01

从表4可知,灰色神经网络预测的精度在传统的 GM(1,1)模型基础上精度提高了很多,灰色神经网络组合和集成预测又在灰色神经网络本身精度很高的情况下又推进了许多.并在灰色神经网络的建模过程中,发现数据修正可提高精度,但不能将全部数据严格修改成某一趋势值,只能修正偏离系统发展趋势较大的数据,即只能修正较异常的数据,否则,若全部数据修正成严格按某种趋势变化,会使"趋势性"平稳,滞后现象增加,预测结果不灵敏,不能较快地反映数据变动的趋势,反而会掩盖系统本身的最新发展规律,导致预测精度变差.因为任一时间序列既有一定的趋势性,又有一定的随机性和动态发展,而这种动态发展又可能形成新的趋势性,由此造成了时间序列数据的复杂发展变化,因此修正数据序列时,既要修正偏离原趋势太大的异常数据,使数据序列的发展变化保持一定的连续性,又要保留偏离趋势值不大的原始数据,否则不能反映数据序列的动态变化和系统新的发展趋势,因此必须在趋势性和预测的灵敏性间取得一个平衡.可根据实际情况定超过某一趋势值的百分之几的数据为异常数据和要修正的数据,即修正数据要把握一个度.

4 结束语

提出了基于失真数据修正的改进型灰色神经网络预测模型和算法,并以南昌火车站旅客发送量预测为例,验证了模型及算法的有效性,收到了很好的预测效果,为这类时间序列的预测提出了可行性途径;在灰色神经网络的基础上提出了灰色神经网络组合与集成预测,并用实例与模型对比说明数据修正七改进型灰色系统、改进型灰色神经网络七灰色神经网络集成可提高预测精度.在数据修正灰色神经网络建模过程中,发现数据修正要在趋势性数据与原始数据间取得平衡,既要修正原始序列中特别异常的数据,又要保留偏离趋势值不大的原始数据而不能全部改为趋势值,可根据实际情况确定异常数据和要修正数据.

参考文献:

[1]陈泽淮,张 尧,武志刚. RBF 神经网络在中长期负荷预测中的应用[J]. 电力系统及其自动化学报,2006,18(1):15-19.
CHEN Zehuai, ZHAGN Yao, WU Zhigang. Application of RBF neural network in medium and long-term load forecasting[J]. Proceedings of the Chinese Society of Universities,2006,18(1):15-19.

[2]中国国家统计局. 中国统计年鉴 2005[M]. 北京:中国统计出版社,2005.

[3]邓聚龙. 灰色系统理论教程[M]. 武汉:华中理工大学出版社,1990.

[4]王翠茹,孙辰军,杨 静,冯海迅. 改进残差灰色预测模型在负荷预测中的应用[J]. 电力系统及其自动化学报,2006,18(1):86-89.
WANG Cuiru, SUN Chenjun, YANG Jing, FENG Haixun. Application of modified residual error gray prediction model in power load forecasting[J]. Proceedings of the Chinese Society of Universities,2006,18(1):86-89.

[5]祖恩三. 云南 GDP 的灰色预测和分析[J]. 经济师,2006(6):272-274.
ZU Ensan. A grey forecasting model and its application in GDP forecasting of yunnan[J]. China Economist,2006(6):272-274.

[6]张大海,江世芳,史开泉. 灰色预测公式的理论缺陷及改进[J]. 系统工程理论与实践,2002(8):140-142.
ZHANG Dahai, JIANG Shifang, SHI Kaiquan. Theoretical defect of grey prediction formula and its improvement[J]. Systems Engineering-Theory & Practice,2002(8):140-142.

作者简介:

严修红,男,1974年生,一级教师,主要研究方向为智能预测.
E-mail:yxh3o9@163.com.

许伦辉,男,1965年生,教授,主要研究方向为智能交通系统、交通环境与交通安全、交通系统建模与仿真,发表论文50多篇.

董世畅,男,1968年生,一级教师,主要研究方向为中学物理.