

上呼吸道感染月发病率的主成分神经网络预测模型

万淑慧¹, 田富鹏²

1. 西北民族大学计算机科学与信息工程学院, 兰州 730030

2. 西北民族大学网络信息中心, 兰州 730030

摘要 在医学卫生领域, 疾病受许多因素的影响, 很难用结构式因果模型解释, 根据神经网络预测是一种行之有效的方法。径向基函数(RBF)神经网络应用于疾病的月发病人数预测时, 由于影响它的气象因素, 如月平均气压、月平均气温、月平均相对湿度、月平均风速、月平均降水量等本身具有很大的相关性, 且维数较高, RBF神经网络的预测精度会下降。针对这一问题, 本文提出了利用主成分分析方法对原输入空间进行重构, 并根据各主成分的贡献率确定网络结构, 从而有效地解决了预测精度下降问题。最后, 以2001年8月至2006年9月甘肃省武威市上呼吸道感染月发病人数的资料验证该方法的有效性。至此, 应该充分考虑人在各时间段的发病特征, 以便更有重点地进行健康防治工作, 有效地降低支气管肺炎对人类的危害, 保障人类的生活品质。

关键词 径向基神经网络; 预测模型; 主成分分析; 仿真; Matlab 软件

中图分类号 R562

文献标识码 A

文章编号 1000-7857(2010)03-0087-03

Principal Component Neural Network Prediction Model for Incidence of Upper Respiratory Tract Infection

WAN Shuhui¹, TIAN Fupeng²

1. Faculty of Computer Science and Information Engineering, Northwest University for Nationalities, Lanzhou 730030, China

2. Network Information Center, Northwest University for Nationalities, Lanzhou 730030, China

Abstract In the medical field, due to the fact that diseases are often affected by many factors, it is difficult to use a structural causal model, while on the other hand, it would be effective to establish a dynamic model, based on their own time-series changes. To predict the number and incidence of diseases, because meteorological factors, the monthly average atmospheric pressure, monthly mean temperature, monthly mean relative humidity, monthly average wind speed, the monthly average precipitation are strongly correlated between themselves and with very high dimensions, the accuracy of the Radial Basis Function (RBF) neural network prediction may be very low. To solve this problem, this paper proposes the use of the Principal Component Analysis (PCA) method to reconstruct the original input space and, based on the contribution rate of all main components to determine the network structure, which will effectively solve the problem of low prediction accuracy. The incidence data of upper respiratory tract infection, from August 2001 to September 2006, in Wuwei City of Gansu Province are used to validate the method. The clinical characteristics in each time period should be duly considered in order to carry out a more focused health prevention and treatment and to effectively reduce the hazards to human bronchial pneumonia.

Keywords radial basis function neural network; prediction model; principal component analysis; simulation; software Matlab

收稿日期: 2009-10-23

基金项目: 国家自然科学基金项目(60673192)

作者简介: 万淑慧, 硕士研究生, 研究方向为数学模型, 电子信箱: shuhuiwan@163.com; 田富鹏(通信作者), 教授, 研究方向为数据挖掘, 电子信箱: tfp@xbmu.edu.cn

0 引言

气象因素是甘肃省武威市上呼吸道感染发病人数预测模型中一种非线性、多变量时间序列。对于非线性系统,一般采用基于非线性自回归滑动平均模型进行预测,但很难为这种模型找到恰当的参数估计方法,因此传统的非线性系统的识别,在理论研究和实际应用方面都存在极大困难。

相比之下,神经网络在这方面显示了明显的优越性,它的优点在于具有模拟多变量而不需要对输入变量作复杂的相关假设,不要求知道输入输出变量间的函数关系,只需通过对输入输出数据训练,获得输入、输出之间的映射关系,从而进行预测。目前,已经有一些比较成熟的神经网络软件包,其中 Matlab 软件的神经网络工具箱应用最为广泛。

1 模型简介

设多重非线性时间序列为

$$x_1(1), x_1(2), \dots, x_1(N)$$

$$x_2(1), x_2(2), \dots, x_2(N)$$

$$x_n(1), x_n(2), \dots, x_n(N)$$

其中, N 为时间序列的长度, n 为序列变量个数。设第 i 个变量的嵌入维数为 D_i , 时间延迟为 $\tau_i, i=1, 2, \dots, n$ 。重构相空间, 得出在相空间有 L 个相空间点: $X_i(1), X_i(2), \dots, X_i(L), L=N-\max_i [(D_i-1)\tau_i]$ 。下面给出具体的建模方法。

1.1 主成分分析方法

主成分分析 (Principal Component Analysis, PCA) [1-3] 的思想是, 多个变量之间往往存在着一定程度的相关性。本文希望通过线性组合的方式, 从这些指标中尽可能地提取信息, 当第 1 个线性组合不能提取更多的信息时, 再考虑第 2 个线性组合……直到所提取的信息与原指标接近为止。

以时间序列 $X_i=[X_i(1), X_i(2), \dots, X_i(L)]_{D_i \times L}$ 为例, D_i 为时间序列 $x_i(1), x_i(2), \dots, x_i(N)$ 的嵌入维数, PCA 具体步骤如下。

1) 标准化样本矩阵 X_i 为 $\tilde{X}_{D_i \times L}$ 。

2) 按下式奇异值分解 $\tilde{X}_{D_i \times L}$:

$$\tilde{X}=U\Sigma V^T \quad (1)$$

$$\text{其中, } \Sigma=\text{diag}[s_1 \ s_2 \ \dots \ s_p \ \dots \ 0 \ \dots \ 0] \quad (2)$$

p 为特征值个数, $s_1 \geq s_2 \geq \dots \geq s_p$ 为对应的奇异值, U 和 V 都是正交矩阵, $U \in F^{D_i \times D_i}, \Sigma \in F^{D_i \times L}, V \in F^{L \times L}$ 。

3) 较大的奇异值作为主要成分被保留, η_{k_i} 为

$$\eta_{k_i} = \sum_{j=1}^{k_i} s_j^2 / \sum_{i=1}^p s_i^2 \quad (3)$$

式中, $k_i=1, 2, \dots$; s_i, s_j 为相应的奇异值; 确定阈值 $0 < \eta_0 < 1$, 如果 $\eta_k > \eta_0$, 则保留前 k_i 个奇异值。

4) 降低变换矩阵 $U_{D_i \times D_i}$ 的维数为 $\tilde{U}_{k_i \times D_i}$ 系统主元为

$$Z_{k_i \times L} \approx U_{k_i \times D_i} X_{D_i \times L} \quad (4)$$

奇异值分解满足正交分解最优性, 与 PCA 具有等价性,

为上述过程提供了理论依据。

1.2 径向神经网络模型

径向神经网络具有良好的非线性逼近能力, 理论上已经证明径向神经网络具有逼近有界闭集上任意连续函数的能力 [4]。

RBF 神经网络 [5-9] 由两层组成, 第 1 层为隐含的径向基层, 第 2 层为输出线形层, 径向基函数是径向对称的, 最常用的是高斯函数:

$$R_i(x) = \exp \left[-\frac{\|x - c_i\|^2}{2\sigma_i^2} \right] \quad i=1, 2, \dots, p \quad (5)$$

其中, x 为 m 维输入向量, c_i 为第 i 个函数的中心, σ_i 为第 i 个感知的变量, p 为感知单元个数, $\|x - c_i\|$ 为 $x - c_i$ 的范数。

输入层实现从 $x \rightarrow R_i(x)$ 的非线性映射, 输出层实现了从 $R_i(x) \rightarrow y_k$ 的线性映射, 即

$$y_k = \sum_{i=1}^q \omega_{ik} R_i(x) \quad k=1, 2, \dots, q \quad (6)$$

其中, q 为输出节点数。

月发病人数预测 RBF 神经网络模型建立步骤如下。

步骤 1 用主成分分析确定 RBF 神经网络的输入, 最后确定为 2 维输入向量和 1 维输出值, 模型输出节点数 $q=1, 56$ 组 2 维输入向量和 1 维输出向量构成训练样本矩阵。

步骤 2 用 newrb 函数设计一个满足一定精度要求的 RBF 网络。格式: $\text{net}=\text{newrb}(x, y, g, s)$, 用 RBF 网络逼近函数时, newrb 函数可自动增加 RBF 网络的隐层神经元, 直到均方误差满足为止。式中, x, y, g, s 分别为输入向量矩阵、目标向量矩阵、均方误差和 RBF 的分布。

步骤 3 用 sim 函数对时间序列进行预测, 格式为 $b=\text{sim}(\text{net}, a)$ 。式中, a, b 分别为待评价时间序列的输入向量和用 RBF 网络对时间序列进行计算的预测值。以上 3 个步骤构成了时间序列预测的 RBF 神经网络月发病人数预测模型。

2 预测结果分析

用主成分方法对影响月发病人数的气象因素: 月平均气压 (X_1), 月平均气温 (X_2), 月平均相对湿度 (X_3), 月平均风速 (X_4), 月平均降水量 (X_5) 进行分析, 分析前先进行数据标准化处理。由因子分析得到表 1, 可以看出, 前两个因子的贡献率达 83.6%, 因此主成分分析取 $m=2$ 作为 RBF 网络模型的输入向量的维数。

计算特征向量矩阵 (表 2), 得到输入向量的表达式为

$$Y_1 = -0.574X_1 + 0.5836X_2 + 0.036X_3 + 0.3654X_4 + 0.4419X_5$$

$$Y_2 = 0.0857X_1 + 0.0841X_2 + 0.727X_3 + 0.4908X_4 + 0.4646X_5$$

取数据的前 56 个作为训练样本, 用 Matlab 7.0 环境下的 RBF 神经网络模型对甘肃省武威市上呼吸道感染后 6 个月发病人数进行预测, 结果见表 3, 网络训练的图形输出结果和预测结果如图 1 所示。

从图 1 和表 3 可知, 经 PCA 预处理输入数据后, RBF 神经网络模型可以较好地预测武威上呼吸道感染月发病率。

表 1 特征根和方差贡献表
Table 1 Total variance contribution

成分序号	初始特征值			提取求和的平方载荷		
	总计	各因子方差贡献率/%	累计方差贡献率/%	总计	各因子方差贡献率/%	累计方差贡献率/%
1	2.502	50.036	50.036	2.502	50.036	50.036
2	1.679	33.572	83.607	1.679	33.572	83.607
3	0.449	8.988	92.596			
4	0.205	4.097	96.693			
5	0.165	3.307	100.000			

表 2 特征向量矩阵
Table 2 Eigenvector matrix

	t_1	t_2
X_1	-0.5740	0.0857
X_2	0.5835	0.0841
X_3	0.0360	0.7270
X_4	0.3654	0.4908
X_5	0.4419	0.4646

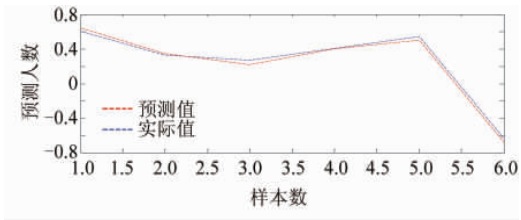


图 1 实际值与预测值
Fig. 1 Actual values and predicted values

表 3 实际值和预测值
Table 3 Actual values and predicted values

	1	2	3	4	5	6
实际值	0.6081	0.3277	0.2745	0.4085	0.5472	-0.6364
预测值	0.6417	0.3467	0.2238	0.4030	0.5016	-0.6815
绝对误差	0.0417	0.019	-0.0507	-0.0055	-0.0456	-0.0451
相对误差/%	6.90	5.80	18.50	1.30	8.30	7.10

3 结论

径向基神经网络预测系统具有良好的性能,在很多领域发挥很大的作用。但是要对某一对象进行较准确的预测,首先要保证训练样本必须是准确的,并且训练样本的数量要足够多,输入向量是目标向量的主要影响因素,本文的不足之处在于,虽然气候因素影响上呼吸道感染,但这并不是此病发病的决定性因素,因此预测误差还是很大,要想使误差减少,必须找到引发上呼吸道感染的决定性因素,作为神经网络的输入预测。

参考文献 (References)

[1] 潘春光, 陈英武, 汪皓. 主成分分析法在基于度量的软件风险评估中的应用[J]. 运筹与管理, 2005, 14(5): 81-84.
Pan Chunguang, Chen Yingwu, Wang Hao. *Logistics and Management*, 2005, 14(5): 81-84.

[2] 向东进. 实用多元统计分析 [M]. 武汉: 中国地质大学出版社, 2005: 120-184.
Xiang Dongjin. *Utility of multivariate statistical analysis* [M]. Wuhan: China University of Geosciences Press, 2005: 120-184.

[3] 朱建平, 殷瑞飞. SPSS 在统计分析中的应用 [M]. 北京: 清华大学出版社, 2007: 171-178.
Zhu Jianping, Yin Ruifei. *SPSS statistical analysis*[M]. Beijing: Tsinghua

University Press, 2007: 171-178.

[4] Boddy A, Bentz E, Thomas M D A, *et al.* An overview and sensitivity study of a multimechanistic chloride transport model [J]. *Cement and Concrete Research*, 1999, 29(6): 827-837.

[5] 焦李成. 神经网络计算 [M]. 西安: 西安电子科技大学出版社, 1993: 85-102.
Jiao Licheng. *Neural network computation* [M]. Xi'an: Xidian University Press, 1993: 85-102.

[6] 许东, 吴铮. 基于 Matlab 6.x 的系统分析与设计——神经网络 [M]. 西安: 西安电子科技大学出版社, 2002: 25-26, 194-199.
Xu Dong, Wu Zheng. *Matlab 6.x -based analysis and design—Neural networks*[M]. Xi'an: Xidian University Press, 2002: 25-26, 194-199.

[7] 飞思科技产品研发中心. 神经网络理论与 Matlab 7 实现[M]. 北京: 电子工业出版社, 2005: 79-81, 116-126.
FECIT Sci -Tech Product R&D Center. *Neural network theory and Matlab 7 realization*[M]. Beijing: Electronic Industry Press, 2005: 44-52, 99-108.

[8] 李季涛, 杨俊峰. 基于径向基神经网络的大连站客运量预测 [J]. 大连交通大学学报, 2007, 28(1): 32-34.
Li Jitao, Yang Junfeng. *Journal of Dalian Jiaotong University*, 2007, 28 (1): 32-34.

[9] 姚亚夫, 彭昊. 一种基于径向基神经网络的组合预测模型 [J]. 郑州大学学报: 工学版, 2008, 29(3): 137-140.
Yao Yafu, Peng Hao. *Journal of Zhengzhou University: Engineering Science Edition*, 2008: 29(3): 137-140.

(责任编辑 朱宇)