文章编号:1006-7639(2012)-01-0130-06

数据挖掘技术在精细化温度预报中的应用

段文广[1,2],周晓军[1],石永炜[3]

(1. 甘肃省兰州市气象局,甘肃 兰州 730020;2. 兰州大学信息学院,甘肃 兰州 730020;
3. 甘肃省兰州市人工影响天气办公室,甘肃 兰州 730020)

摘 要:简要介绍了精细化天气预报和气象数据挖掘应用的现状,在对BP神经网络预测方法详细分析的基础上,研究了基于时间序列数据挖掘实现精细化温度预报的方法。该方法基于时序分析技术,建立起适合于BP神经网络的输入样本模型,通过反复学习从温度时序中建立预测模型,将其用于未来24 h的精细化温度预报。同时,对BP神经网络算法和步骤做了简要介绍,针对原有的BP算法存在的不足,做了一些改进。最后,通过对预测挖掘系统的设计和在Matlab6.5仿真平台上的试验,建立了温度预报模型,以兰州市观测站数据为时间序列研究对象,对精细化温度预报进行了仿真实现。对基于时序的数据挖掘理论的应用和开发精细化温度预报方法做了有益的探索。

关键词:数据挖掘;精细化温度预报;BP神经网络;预测模型
中图分类号:P456.9 **文献标识码**:A

引 言

天气预报的精细化,是天气预报技术发展到相对成熟阶段的必然趋势,也是目前气象服务面临的迫切需求,精细化天气预报是当今国际气象科学发展的趋势,美国、日本等发达国家凭借技术优势,可以提供时间间隔为1 h,空间分辨率达5 km的精细天气预报。国内技术条件的限制,精细天气预报业务化工作才刚刚起步,目前国内的精细化天气预报一般利用非静力平衡中尺度数值预报模式(MM5模式),以北京T213数值预报产品为初始场,引进本地地面和探空实时资料,实现本地区域中尺度精细化天气预报,预报分辨率为20 km×20 km。如胡文东等宁夏精细化预报产品显示与评价业务系统[1],魏秀兰等菏泽市温度精细化预报[2]。但是对于空间分辨率低于5 km的乡镇级城乡精细化预报来说,就无能为力了。

数据挖掘的应用非常广泛,但很少有涉及到气象资料的专业数据挖掘系统。随着数据挖掘技术在气象系统应用的普及,许多专家认识到数据挖掘的应用前景,并从建立气象数据库、气象预报、气候预测等不同角度的应用方面进行了研究,主要的研究和应用有通过建立干旱指标挖掘系统对干旱进行逐日预测[3];基于项目序列的空间关联规则挖掘算法[4];采用数据挖掘技术中的关联规则从气象历史数据库中发现灾害范例,建立范例库,指导灾害预报的准确性[5]等。

兰州市区域气象观测网主要由3个国家级台站,90个区域气象自动站构成,提供区域性高时空分辨率的中小尺度灾害性天气观测数据,应用于短时临近预报服务系统,为开展城乡精细化预报打下了基础。图1所示为兰州市区域站分布图,基本实现了所有乡镇的布点。到目前为止,兰州市区域气象观测网形成的近5 a的历史数据已经达到1千多万个气象要素值。预报人员面对大量的数据资料缺乏有效的工具去提取有价值的信息来指导中小尺度的精细化天气预报和短时临近预报预警业务,仅仅是依据当前区域站的实时数据,根据预报人员的经验来做一些订正预报。

利用数据挖掘技术对区域气象观测网温度数据进行探索,挖掘一些有用的模式,拓展温度预报方法,提高温度要素的精细化预报质量和区域气象观测网资料的利用效率是本文研究的重点。

收稿日期:2011-08-17;**改回日期**:2011-12-22
作者简介:段文广(1975-),男,汉族,甘肃酒泉人,工程师,计算机应用工程硕士,研究方向为数据挖掘. E-mail:ymdcr@163.com

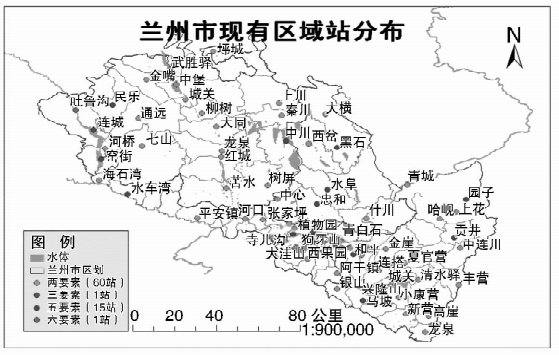

图 1　兰州市区域站分布图

Fig. 1　The distribution of Lanzhou regional weather stations

1　时间序列的数据挖掘方法

基于一个或多个时间序列的数据挖掘,它可以从时序中抽取时序内部的规律用于时序的数值、周期、趋势分析和预测等。一般来说,时间序列数据挖掘的目标有 2 个:一是时间序列建模,即洞察产生时间序列的机制或根本因素;二是时间序列预测,即预测时间序列变量的未来值。主要有趋势分析、周期性分析、序列模式分析、相似性分析等几个研究方面[8]。将数据挖掘的思想引入到时间序列数据分析中,对时间序列数据进行挖掘,从中发现蕴含的系统规律,将其用于时间序列系统的分析和预测,这将很好地弥补传统时序分析方法的不足,为时序问题的研究提供了一种新的思路和方法。

本文中,时间序列数据挖掘的目标是预测,即预测时间序列变量的未来值。首先采取了时间序列分析中的建模理论建立起适合实际应用的准确模型,再利用数据挖掘理论中 BP 神经网络技术来实现对未来一天的温度进行预报,解决了有关的实现技术等一系列关键问题。本文研究的方法针对气象要素如温度进行预报,这种方法对于拓展天气预报方法,提高预报精度具有很好的实际意义。

2　人工神经网络预报方法

人工神经网络(Artificial Neural Networks,简称 ANN)系统由于具有信息的分布存储、并行处理以及自学习能力等优点,已经得到广泛应用。尤其是基于误差反向传播(Backproragation)算法的多层前馈网络(Multiple - Layer Feedforward network,简称 BP 网络),可以以任意精度逼近任意的连续函数[8],所以广泛应用于非线性建模、函数逼近、模式分类等方面。人工神经网络具有高度的并行性、高度的非线形全局性、良好的容错性与联想记忆功能、十分强的自适应、自学习功能,能够实现复杂的逻辑操作和非线形关系系统[9]。神经网络的处理过程主要是通过网络的学习功能找到一个恰当的连接加权值来得到最佳结果。其比较典型的学习方法是回溯法(Back - propagation)。它通过将输出结果同一些已知值进行一系列比较,加权值不断调整,得到一个新的输出值,再经过不断的学习过程,最后该神经网络得到一个稳定的结果[10]。

3　基于时间序列数据挖掘实现精细化温度预报的方法

基于神经网络可进行温度预报的依据:人工神经网络具有可以以任意精度逼进任意非线性过程的特性,由此能模拟温度的变化规律,其算法简单,计算速度快,预测误差小。本文将数据挖掘理论中的神经网络方法与时间序列分析理论相结合,将其应用于精细化温度预报中,建立起基于神经网络的时间序列模型,以达到能够提高预报精度的目的。利用人工神经网络的学习功能,用大量的历史样本对其进行训练,当网络训练完成,此网络即可作为非线性预测器。对于新的时间序列值,经过计算即可得出预测值。

3.1　BP 神经网络算法简介

误差反向传播训练算法,即 Error Back Propagation Training,简称 BP 算法,本文以 3 层 BP 神经网络为研究对象,其一般结构如图 2 所示。网络由输入层、输出层和隐藏层组成。

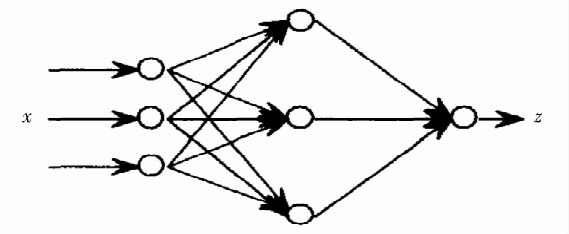

图 2　BP 神经网络结构图

Fig. 2　The structure of BP neural networks

BP 神经网络工作时,学习过程的信号正向传播时,输入样本从输入层传入,若实际输出与期望输出

不符,则转入误差的反向传播阶段。误差反传时将输出误差向输入层反传,获得各层单元的误差信号并将其作为修正各单元权值和阈值的依据。这种信号正向传播与误差反向传播是循环进行的。权值和阈值的不断调整过程就是网络的训练过程。此过程一直进行到网络输出的误差减少到预先设定的值为止。

3.2 BP 神经网络算法的具体实现步骤

1. 对 BP 神经网络进行初始化

(1)初始化参数

选定一个结构合理的神经网络,确定网络层数和各层单元数;

确定输入样本值 $X(x_1,x_2,\cdots,x_i,\cdots,x_n)$;

确定目标输出值 $C_k(c_1^k,c_2^k,\cdots,c_q^k)$;

确定网络最大允许误差,也就是网络精度 $E_{\min}$;

(2)将各输入样本值提供给网络输入层的各神经元。

2. 向前的传播阶段

(1)计算网络隐藏层各神经元的输入(激活值) S_j,也就是输入层各单元的输出。其中:

$$S_j = \sum_{i=1}^{n} W_{ij}x_i - \theta_j \tag{1}$$

(2)计算输出层各神经元的响应(即输出值) y_t:

$$y_t = f(l_t) = \frac{1}{1+e^{-l_t}} \tag{2}$$

在此阶段,信息从输入层经过逐级的变换传送到输出层。这个过程也是网络在完成训练后正常运行时执行的过程。

3. 向后传播阶段

(1)计算隐藏层与输出层之间的新权值 $V_{jt}(N+1)$ 和新阈值 $\gamma_t(N+1)$,N 为学习次数,$d^k_{t(t=1,2,\cdots,q)}$ 为输出层各神经元的校正误差:

$$V_{jt}(N+1) = V_{jt}(N) + \eta d_t^k b_j \tag{3}$$

$$\gamma_t(N+1) = \gamma_t(N) + \eta d_t^k \tag{4}$$

(2)计算输入层与隐藏层之间的新权值 $W_{ij}(N+1)$ 和新阈值 $\theta_j(N+1)$,$e^k_{j(j=1,2,\cdots,p)}$ 为隐藏层校正误差:

$$W_{ij}(N+1) = W_{ij}(N) + \eta e_j^k x_i \tag{5}$$

$$\theta_j(N+1) = \theta_j(N) + \eta e_j^k \tag{6}$$

4. 随机从输入样本中再取一组样本或再取原样本值重复 2、3 过程,直至网络全局误差小于预先设定的限定值 $E_{\min}$(网络收敛)。

3.3 BP 算法存在局限性及改进方法

尽管 BP 反向传播算法应用很广泛,但是它同样也存在自身一些不足的地方,存在着比如收敛速度慢、容易陷入局部极小等问题[11],其主要表现在其训练过程的不确定上。针对收敛速度慢及函数陷入局部极小值的改进方法结合本文要求,采用增加动量因子和自适应学习速率 2 种改进方法。

4 基于温度序列的 BP 神经网络预报挖掘系统的设计

4.1 BP 神经网络结构设计

如前文所提到的,在进行 BP 网络的设计时,一般从网络的层数、每层中单元数的个数和激活函数、初始值以及学习速率、样本数量等几个方面来进行考虑。

(1)BP 神经网络层数的确定

通常前馈神经网络采用的是 3 层网络结构,本文就以 3 层结构的神经网络为讨论对象,即网络有一个输入层、一个隐含层和一个输出层。

(2)输入层和输出层单元数的确定

根据某气象台站历史的 24 h 温度资料,预测未来一天的 24 h 温度。所以设计输入层单元为 24 个,输出层单元也为 24 个。

(3)隐含层单元数的确定

以经验值来确定隐含层的单元数:网络的输入层有 N 个单元,隐含层有 $2N+1$ 个单元。由上文中确定输入层 $N=24$。由此,可确定隐含层的单元数为 49 个。

(4)初始权值的选取

初始权值的选取是非线性的,初始值对于学习是否达到局部最小、是否能够收敛以及训练时间的长短关系很大。所以,一般取初始权值在(-1,1)之间的随机数。

(5)学习速率

学习速率决定每一次循环训练中所产生的权值

变化量。一般倾向以选取较小的学习速率以保证系统的稳定性。学习速率选取的范围在 0.01 ~0.8 之间。

(6)样本数量的确定

对于某气象台站观察的温度数据,一般具有几十年的历史资料,对于训练样本数量的选择应该有一个合适的范围,观察这些温度数据,具有明显的日特性和季节特性,因此选取适当的样本能使神经网络既能学习到数据的基本周期性规律,又不至于学习不足或者过分学习。具有季节特性的时间序列中的数据会同那些领先或滞后 12 个月的相应数据存在某种程度的相关。因此采用的样本为相关性较强的本年前 2 个月的数据和前几年与本季度相对应的 3 个月的数据。

4.2　对输入、输出数据的预处理

对于激活函数 Sigmoid 函数来说,它的性质要求输入样本必须保证在正常范围内的[-1,1]区间,必须对输入输出数据进行预处理,大多采用归一化预处理方法。即找出历史样本数据中的最大值和最小值,用当前值减去最小值后除以最大值与最小值的差,即线性函数转换,表达式如下:

$$y = \frac{(x - x_{\min})}{(x_{\max} - x_{\min})} \tag{7}$$

其中:x、y 分别为转换前、后的值,$x_{\max}$、$x_{\min}$分别为样本的最大值和最小值。同样,当网络学习完成以后,必须按与输入相反规则进行变换以求出具体的数据。

4.3　温度序列的仿真试验

选择的仿真平台是 MathWorks 公司开发的 Matlab6.5。

首先取相关性较高的本季度的日 24 h 温度历史数据,利用这些历史数据,通过仿真试验来预测未来一天的 24 h 温度,最后与实际的 24 h 温度值进行对比。以下是具体的实现过程:

(1)权值和阈值初始化函数

函数 INITFF 对 BP 网络的权值和阈值赋初始值,网络层数为 3 层,调用格式为:

[*W*1,*B*1,…] = *initff*(*P*, *Si* ‘*F*1’,…, *Pn*, *Sn*, ‘*F*’)

其中,*P* 为 *R* ×24 网络输入矩阵,*R* 为网络输入节点数;*Si* 为第 *i* 层节点数;*Fi* 为第 *i* 层节点的传递函数;*Wi* 为第 *i* 层权矩阵;*Bi* 为第 *i* 层的阈值矢量。

(2)网络训练函数

网络训练函数是一个直接用于循环训练一个 BP 神经网络终达到允许目标误差的函数。MATLAB 神经网络工具箱提供的 BP 神经网络训练函数有 *Traingd*、*Trainbp*、*Trainbpx*、*Trainlm* 和 *Trainbr* 等[12]。这里采用函数 *trainbpx*( )来训练网络。它的调用格式为:

[*W*1,*B*1,*W*2,*B*2,…*TE*,*TR*] = *trainbpx*(*W*1,*B*1, *F*1,…,*P*,*T*,*TP*)

其中,*Wi* 为第 *i* 层权矩阵,*i* = 1,2…;*Bi* 为第 *i* 层的阈值矢量,*i* = 1,2…;*Fi* 为第 *i* 层节点的传递函数,*i* = 1,2;*P* 为 *R* × *Q* 输入矩阵;*T* 为 *Q* 目标矩阵;*TP* 为训练参数。

(3)网络仿真函数

函数 *SIMUFF* 用来计算 3 层以内 BP 神经网络在给定输入下的输出,调用格式为:

[*A*1,*A*2,…,*AN*] = *simuff*(*P*,*W*1,*B*1,‘*F*1’,…, *WN*,*BN*,‘*FN*’)

其中,*P* 为网络输入矢量;*Wi* 为第 *i* 层权矩阵;*Bi* 为第 *i* 层阈值;*Fi* 为第 *i* 层节点传递函数;*Ai* 为第 *i* 层节点输出。图 3 为在 Matlab6.5 环境下某次温度序列仿真运算的结果曲线图,可以看模拟的曲线与实际值大致是拟和的。

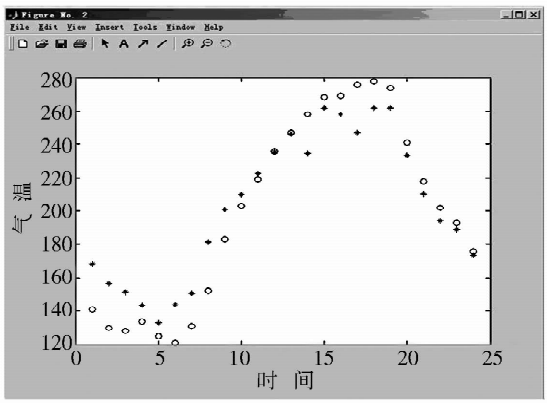

图 3　温度序列仿真运算结果

Fig. 3　Simulation results of temperature series

(4)仿真实例

选择预测 2009 年 7 月 22 日 08 时至 7 月 23 日 07 时的温度预报值,经过多次仿真试验,选取 2005 ~2008 年逐时次的温度值作为训练值,采用函数 *trainbpx*( )来训练网络。选取兰州观测站 2009 年 5 月、6 月和 7 月 22 日 08 时前的数据和 2005 ~2008 年 6、7、8、9 月的数据进行仿真计算,得到温度预报

值。经仿真试验后得到的 2009 年 7 月 22 日 08 时至 7 月 23 日 07 时的温度预报结果见表 1，从表中可以看出大多值误差都不是很大。

表 1 兰州观测站 2009 年 7 月 22 日的温度预报仿真试验结果

Tab. 1 Simulation results of temperature forecast at Lanzhou observational station on July 22, 2009

时次	实际温度值/°C	预测温度值/°C	相对误差/%	平均相对误差/%
072208	19.3	17.9	-7.25	-5.47
072209	20.7	19.8	-4.35	-5.47
072210	21.2	22.2	4.72	-5.47
072211	24.2	23.6	-2.48	-5.47
072212	25.8	23.7	-8.14	-5.47
072213	27.4	25.0	-8.76	-5.47
072214	28.9	26.9	-6.92	-5.47
072215	31.0	29.9	-3.55	-5.47
072216	29.9	27.6	-7.70	-5.47
072217	28.2	28.0	-0.71	-5.47
072218	27.5	28.1	2.18	-5.47
072219	26.6	27.2	2.26	-5.47
072220	26.0	25.8	-0.77	-5.47
072221	25.5	24.5	-3.92	-5.47
072222	25.0	23.3	-6.80	-5.47
072223	24.2	22.0	-9.09	-5.47
072300	23.7	20.8	-12.24	-5.47
072301	22.8	20.6	-9.65	-5.47
072302	22.3	20.0	-10.31	-5.47
072303	22.3	19.3	-13.45	-5.47
072304	20.9	19.3	-7.66	-5.47
072305	19.4	18.5	-4.64	-5.47
072306	19.4	18.2	-6.19	-5.47
072307	19.1	18.0	-5.76	-5.47

通过大量的仿真试验结果表明，基于本文所建模型的研究算法在无转折性天气过程时，如图 4 温度预测值与实际值比较所示，预报误差大多在 10% 以内。

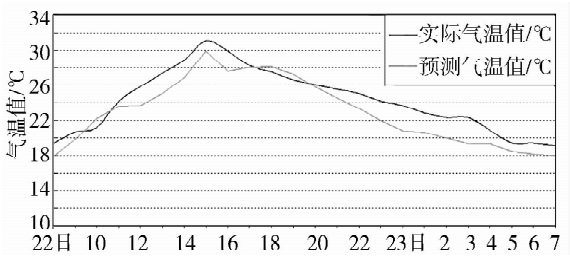

图 4 2009 年 7 月 22 日 08 时至 23 日 07 时气温预测与实际值比较

Fig. 4 Comparison between true temperature and forecast value

5 总 结

选择了气象数据中的极小的一部分（区域站温度数据）进行数据挖掘尝试，重点研究了 BP 神经网络数据挖掘中的若干技术和常用算法。通过研究和分析，对时间序列的神经网络挖掘技术应用到气象数据中，从一个新的角度对气象数据进行处理，并使用上述算法进行气象数据时间序列挖掘，较之使用传统的统计方法，挖掘速度和预报值均有较大的提高，解决长期以来一直困扰预报员的“面对堆积如山的数据无从下手，只好置之不理”的局面，结合其他的预报模式和方法，有利于预报模式的不断改进，最终产生较为理想的预报模式，这将为气象数据分析提供一种新的实用的方法。

参考文献：

[1] 胡文东，丁建军. 宁夏精细化预报产品显示与评价业务系统[J]. 气象科技，2004. 32(5)：367 - 371.
[2] 魏秀兰，候艳丽，等. 菏泽市气温精细化预报[J]. 山东气象，2004. 24(2)：45.
[3] 王红霞，朱喜林. 气象数据仓库及其上数据统计和挖掘[J]. 太原理工大学学报，2006(增刊)：48 - 51.
[4] 何靖，王丽珍，邹力鹃. 基于云南气象数据的空间关联规则挖掘[J]. 计算机工程与应用，2003，34：187 - 190.
[5] 赵鹏，倪志伟，贾兆红. 利用数据挖掘技术从气象数据库中建立范例库[J]. 微机发展，2002，3：67 - 70.
[6] 张乃尧，阎平凡. 神经网络结构设计的理论与方法[M]. 北京：清华大学出版社，1998.
[7] 韩立群. 人工神经网络理论、设计及应用(第二版)[M]. 北京：化学工业出版社，2007.
[8] Jiawei Han, Micheline Kamber. 范明，孟小峰译. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2006.
[9] 本社. 人工神经网络原理及应用[M]. 北京：科学出版社，2006.
[10] 吉根林，张立明. 数据挖掘技术及其应用[J]. 南京师大学报，2000(2)：45 - 47.
[11] 陆琼瑜，童学锋. BP 算法改进的研究[J]. 计算机工程设计与研究，2007，28(2)：158 - 160.
[12] BP 神经网络在 MATLAB 上的方便实现[J]. 新疆石油学院学报，1999，11(2)：42 - 46.

Application of Data Mining Technique on Refined Temperature Forecast

DUAN Wenguang[1,2], ZHOU Xiaojun[1], SHI Yongwei[3]

(1. *Lanzhou Meteorological Bureau of Gansu Province*, *Lanzhou* 730020, *China*; 2. *Information College*, *Lanzhou University*, *Lanzhou* 730000, *China*; 3. *Lanzhou Weather Modification Office*, *Lanzhou* 730020, *China*)

Abstract: The paper introduces the domestic and international current situation about the development of refined weather forecast and data mining application. On the basis of detailed analysis of BP neural network forecasting method, this paper researches a data mining method based on time series analysis technology which can be used on refined temperature forecast. This method can build an input sample pattern which is suit for the BP neural networks of data mining and finally establish a predictive model by studying temperature time series again and again, which used for the next 24 hours refined temperature forecast. At the same time, a brief introduction of the algorithm and steps of the BP neural network is given out in the paper, and some further improvement is made aiming at the deficiency of the original BP algorithm. Finally, through design data mining system and test on the Matlab6.5 simulation platform, the temperature forecast model was established. This study had done some helpful exploration on application of data mining theory based on time series analysis technology and developed method of refined weather forecast.

Key words: data mining; refined temperature forecast; BP neural networks; forecast model