基于覆盖算法的大气质量预测

施 尧[1],赵 勇[1],杨雪洁[1],赵 姝[1],张燕平[1],吴有训[2],王克强[2]
(1.安徽大学 智能计算与信号处理教育部重点实验室,安徽 合肥 230039;
2.宣城市气象局,安徽 宣城 242000)

摘 要:提出了一种应用人工神经网络进行大气质量预测的方法,即采用多层前向网络的覆盖算法和时间序列进行短期的空气质量预测。针对空气污染的特点,选取了对空气质量有重要影响的气象因子的相关数据,并对其进行测试。实验结果表明:将该方法应用于空气质量预测,效果良好,学习速度快,识别率高,具有较强的实用价值,为实现大气质量预测提供了一种准确高效的方法。

关键词:覆盖算法;时间序列;空气质量等级;预测
中图分类号:TP183 **文献标识码**:A **文章编号**:1673-629X(2008)07-0190-03

Application of Covering Algorithm to Prediction of Air Quality

SHI Yao[1], ZHAO Yong[1], YANG Xue-jie[1], ZHAO Shu[1]
ZHANG Yan-ping[1], WU You-xun[2], WANG Ke-qiang[2]
(1.Ministry of Edu. Key Lab. of Intelligent Computing & Signal Processing at Anhui Univ., Hefei 230039, China;
2.Xuancheng Meteorological Office, Xuancheng 242000, China)

Abstract: A new method of prediction of air quality is proposed based on neural networks, which is the covering algorithm of multi-layer neural networks and time series. According to the characteristics of the prediction of air quality, selects the data that greatly influence the air quality. The experimental result shows that the performance of prediction of air quality is favorable, the learning speed is fast and the rate of accurate is high, so it has a practical value. It provides a precise and efficient way for the prediction of air quality.

Key words: covering algorithm; time series; air quality standard; prediction

0 引 言

随着现代工业的迅速发展,城市环境污染问题已日趋严重,城市的大气环境质量对人体健康的影响越来越受到人们的普遍关注,因此实现对空气质量的预测有着非常重大的理论意义和实践价值。目前在传统的几种利用人工神经网络进行空气质量预测的方法中[1],被广泛采用的灰色系统预测能给出发展趋势,但它只适用于呈近似指数增长的数据序列,对波动性较强的序列的预测效果较差;多层前向网络BP算法本身存在收敛速度慢、网络容错能力(tolerant capacity)差、算法不完备(容易陷入局部最小)等缺陷。

针对BP算法的缺点,张铃教授和张钹院士提出了多层前向网络的覆盖算法,该算法应用范围很广,可以有效地处理预测等问题。而文中则是尝试利用时间序列对原始数据进行处理,然后利用多层前向网络的覆盖算法来对影响空气质量(文中的大气质量、空气质量均指PM_{10}浓度)的数据进行分类,然后分析、检验覆盖算法在气象预测中的预测效果,期望在实践中找出更易操作、更精确的预测方法,提高空气质量预测业务水平。

收稿日期:2007-10-05
基金项目:国家自然科学基金资助项目(60475017,60675031);973计划资助项目(2004CB318108);安徽省教育厅重点自然科学研究项目(2006KJ015A);安徽省自然科学基金资助项目(0504200208);安徽省教育厅自然科学研究项目(2005kj053);安徽大学211工程学术创新团队
作者简介:施 尧(1983-),男,安徽合肥人,硕士研究生,研究方向为机器学习;张燕平,教授,研究方向为神经网络、数据挖掘、人工智能。

1 时间序列

对于随时间变化的动态数据,在建立动态系统模型时习惯地选择时间进程,即将时间作为基本自变量的分析方法,构成时间序列[2]。时间序列问题作为数据挖掘中的一类重要问题,其重要性表现在现实世界中大量数据的采集与时间相关,数据具有时间上的关联性。利用时间序列模式的数据挖掘,可以得到数据

中蕴含的与时间相关的有用信息,实现知识的提取。时间序列从直观上说,决策属性值的变化直接受到样本的各特征属性的影响,因此直接利用原始数据不仅简单明了,而且容易发现数据瞬时变化的特性。对样本集 $\{x_1,x_2,\cdots,x_n\}$,每一个训练样本 x_i 的输入数据都是按一定的关系及预测目标(决策属性)组成的子集,若定义所有的样本长度相等,即具有相同数目特征属性 $v_{i,j}$ 以及由特征属性所决定的决策属性(类别标记)c_i,那么将其记为 $x_i=(v_{i,1},v_{i,2},\cdots,v_{i,m},c_i)$, $i=1,2,\cdots,n;j=1,2,\cdots,m$($i$ 表示样本数,j 表示特征属性数)。如大气质量预测中的日平均气温、日平均气压、日平均相对湿度、日总云量、日低云量、日降水量、日风速七个相关量影响首要空气污染物 PM_{10}。那么可以得到:

$$x_1=(v_{1,1},v_{1,2},\cdots,v_{1,7},c_1);$$

$$x_2=(v_{2,1},v_{2,2},\cdots,v_{2,7},c_2);$$

……

$$x_n=(v_{n,1},v_{n,2},\cdots,v_{n,7},c_n)$$

这样就形成了按日时间序列,但仅以某一固定时段(例如天)来构成基本的样本向量,并不能有效考虑时序数据的前后关系和相互作用。

为了能恰当地反映时序性质的影响,将前面 k 天的特征属性对当日天气的影响也考虑进来,从而可以得到 k 日时间序列:

$$y_1=(v_{1,1},v_{1,2},\cdots,v_{1,7},v_{2,1},v_{2,2},\cdots,v_{2,7},\cdots,v_{k,1},v_{k,2},\cdots,v_{k,7},c_k);$$

$$y_2=(v_{2,1},v_{2,2},\cdots,v_{2,7},v_{3,1},v_{3,2},\cdots,v_{3,7},\cdots v_{k+1,1}v_{k+1,2},\cdots,v_{k+1,7},c_{k+1});$$

……

$$y_{n-k+1}=(v_{n-k+1,1},v_{n-k+1,2},\cdots,v_{n-k+1,7},v_{n-k+2,1},v_{n-k+2,2},\cdots,v_{n-k+2,7},\cdots,v_{n,1},v_{n,2},\cdots,v_{n,7},c_{n-k+1})$$

这样每个样本就包含了 $7*k+1$ 个属性,提供了更多的可用信息,为后面使用覆盖算法学习提供了很好的学习样本。

2　覆盖算法

2.1　领域覆盖算法

多层前向神经网络的设计就相当于用若干"球形领域"将输入 x^i 按其所属的类把它们划分开来。一个最简单的方法,就是对每一类,用一组球形领域将属于该类的 x^i 覆盖住,又不覆盖不属于该类的 x^i,于是不同类的输入被不同组的球形领域所覆盖,然后再将属于同组球领域对应的神经元的输出用或门集中起来。这种分类器的设计方法称为领域覆盖算法[3,4]。

设样本集 K 为:$K=\{x^1,x^2,\cdots,x^k\}$,$x^i\in R^n$ 将 K 分为 s 个子集 $K^1=\{x^1,x^2,\cdots,x^{m(1)}\},\cdots,K^s=\{x^{m(s-1)+1},x^{m(s-2)+2},\cdots,x^k\}$,现在求一个三层网络,使输入集通过此网络后,属于 K^i 的点输出为 y^i。

令样本的输出为 y^t 的样本标号的集合为 $I(t)$(即 $I(t)=\{i\mid y^i=y^t\}$),其对应的输入集合记为 $P(t)$,$t=0,1,2,\cdots,k-1$。并将输入样本的上标按 $I(0),I(1),\cdots,I(k-1)$ 顺序排序。于是,若能够取一批"球形领域"$\{C_j^t,t=0,1,\cdots,k-1;j=1,2,\cdots,k_t\}$,令 $C^t=\bigcup C_i^t$,使得:C^t 只覆盖 j 属 $I(t)$ 的 x^j,而不覆盖 j 不属于 $I(t)$ 的 x^j,且 C^t 互不相交。

任取尚未被覆盖的点 x^i,令 $P(t)=\{x^i\mid i\in I(t)\}$,对样本输入 $x^i\in P(t)$。令

$$d^1(i)\max_{j\in I(t)}\{<x^i,x^j>\}$$

$$d^2(i)\max_{j\in I(t)}\{<x^i,x^j>>d^1(i)\}$$

$$d(i)=\frac{d^1(i)+d^2(i)}{2}$$

$$\theta_i=d(i),W=(w^t),\theta=(\theta_i)\ i=1,2,\cdots,k$$

其中 $<x,y>$ 表示 x,y 的内积。

2.2　领域覆盖算法的步骤

领域覆盖算法(已知样本集 $K=\{r^t=(x^t,y^t),t=0,1,2,\cdots,p-1\}$,且每个 x^t 的长度相等):

(1) 求各 y^t 对应的领域集 $D(t)$,$t=0,1,2,\cdots,k-1$(k 为样本的不同输出的个数)。

(2) 按某种求最小(次小)覆盖的算法,从 $D(t)$ 中求出 $P(t)$ 的最小(或次小)覆盖 $D'(t)$,$t=0,1,2,\cdots,k-1$。

(3) 令 $I'(t)$ 是 $D'(t)$ 中领域对应的指标集,对 $D'(t)$ 中的领域 $D_i(t)$,做对应的神经元 $A_i^t,i\in\bigcup I'(t)$。

(4) 构造网络:第1元件层,由元件 $A_i^t(i\in\bigcup I'(t),t=0,1,2,\cdots,k-1;i=1,2,\cdots,k_t)$ 构成。

(5) 若输出是实值向量时,增加一个隐含层,其元件设为 O_t,O_t 是或门,其输入为 $A_i^t(i\in I'(t))$ 的输出。不然转入(6)。

(6) 输出层,取 m(m 是输出向量的维数)个神经元 $B_s,s=1,2,\cdots,m$。

若输出是实值向量时,B_s 取为线性元件,其对应函数为

$$f(x)=\sum_{t=0}^{k-1}y_s^tO_t(A^t(x)),s=1,2,\cdots,m$$

注:单独符号 A_i,O_i 等表示神经元,而符号 $A_i(x),O_i(x)$ 表示对应的神经元的功能函数。其中 $A^t(x)=(A_1^t(x),A_2^t(x),\cdots,A_{k_t}^t(x))$ 表示一向量;

$O_t(A^t(x))$ 表示 $O_i(x)$ 的输入是 $A^t(x)$ 时输出的值。$i = 1,2,\cdots,k$；B_s 与 O_t 的连接权为 y_s^t。

3 利用覆盖算法对大气质量进行预测

3.1 应用思路

用覆盖算法实际上就是对已有的气象因子的数据进行学习，实现分类器的功能，然后对后来的特征属性的数据进行分析，将其与已分类的类别进行比较，将其归类，这样就完成了预测[5~7]。

3.2 数据处理

文中采用宣城市气象局气象观测资料、宣城市环境保护监测中心提供 PM_{10} 浓度资料(从 2003 年 1 月 1 日到 2005 年 12 月 31 日共 1096 天)。每条数据包括：日平均气压、日平均气温等(日平均相对湿度、日平均总云量、日平均低云量、日降水量、日平均风速)7 个特征属性以及影响空气质量的决策属性 PM_{10}(浓度值)。其中 PM_{10} 为首要污染物。

在进行覆盖之前，要对数据进行加工处理，其步骤如下：

(1) 将原始数据中的决策属性 PM_{10} 的值按照国家标准 GB3095－1996(见表 1)进行分类，相应的空气质量等级确定方法见表 2(标准来源于 http://www.instrument.com.cn/bbs/shtml/20070131/732193/.)。

表 1 空气污染指数对应的污染物浓度限制

污染指数	污染物浓度(mg/m³)				
API	SO_2 (日均值)	NO_2 (日均值)	PM_{10} (日均值)	CO (小时均值)	O_3 (小时均值)
50	0.050	0.080	0.050	5	0.120
100	0.150	0.120	0.150	10	0.200
200	0.800	0.280	0.350	60	0.400
300	1.600	0.565	0.420	90	0.800
400	2.100	0.750	0.500	120	1.000
500	2.620	0.940	0.600	150	1.200

表 2 空气污染指数范围及相应的空气质量类别

空气污染指数 API	空气质量状况
0~50	优
51~100	良
101~200	轻度污染
201~300	中度污染
>300	重污染

(2) 将原始数据分为春夏秋冬 4 种季度模型。

(3) 利用时间序列技术将季度模型的数据处理成时间序列。

(4) 选取一定量的样本作为学习样本和测试样本。

3.3 实验结果

以冬季模型为例，首先以按日时间序列进行处理(其中：2003 年 1 月 2 月 12 月，2004 年 1 月 2 月 12 月，2005 年 1 月共 212 个样本作为学习样本，预测 2005 年 2 月的 PM_{10} 等级)，结果见表 3。

表 3 按日时间序列测试结果

学习样本数	测试样本数	识别率%	覆盖数	拒识个数	训练时间	测试时间
212	28	62.9630	115	10	0.1250	0

通过测试发现按日时间序列进行学习预测的结果不是很理想：识别率比较低，覆盖数比较多，即所需的网络元件数较多，拒识数也较多，没能达到预想的效果。因此考虑到空气质量是无时无刻不在变化的，而且带有明显的时序性，空气质量状况不仅受到当日的污染因素的影响，还会受到前面几日的大气质量状况的影响，所以在选取时间序列的时候选取了 5 日交叉时间序列(其中：2003 年 1 月 2 月 12 月，2004 年 1 月 2 月 12 月，2005 年 1 月 2 月共 228 个样本作为学习样本，预测 2005 年 12 月)，结果见表 4。

表 4 5 日交叉时间序列测试结果

学习样本数	测试样本数	识别率%	覆盖数	拒识个数	训练时间	测试时间
228	27	90.2439	69	14	0.0780	0

测试表明：识别率有了明显的提高。可见利用时间序列模式进行数据挖掘，可以得到数据中蕴含的与时间相关的有用信息，实现了知识的提取，提高了预测的准确性。因而想到如果增加时间序列的长度会不会获得更好的效果，所以将时间序列的属性增多，选取了 10 日交叉时间序列(其中：2003 年 1 月 2 月 12 月，2004 年 1 月 2 月 12 月，2005 年 1 月 2 月共 228 个样本作为学习样本，预测 2005 年 12 月)，结果见表 5。

表 5 10 日交叉时间序列测试结果

学习样本数	测试样本数	识别率%	覆盖数	拒识个数	训练时间	测试时间
213	22	100	66	2	0.0780	0

通过图 1 可以发现，用 10 日时间序列进行学习测试的效果非常理想，对比按日时间序列和 5 日时间序列，识别率达到了 100%，而且覆盖数也减少很多，拒识数也有所减少，达到了预期的效果。

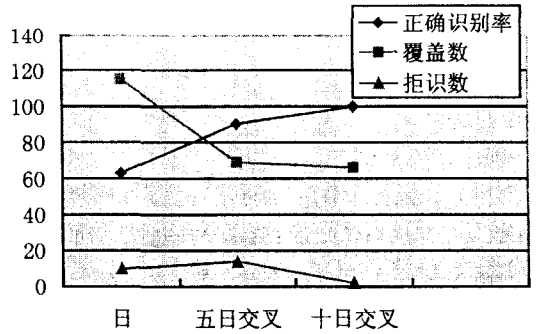

图 1 3 种时间序列的性能比较

(下转第 196 页)

航拍机场跑道这样一类特定目标的识别之上，是非常有效的。

图 7 定位出另一跑道位置

参考文献：

[1] 叶 斌，彭嘉雄. 基于结构特征的军用机场识别与理解[J]. 华中科技大学学报，2001，29(3)：39－42.

[2] LUO Jun，YANG Wei－ping，SHEN Zhen－kang. Automatic target recognition of airfield runway in infrared images[J]. Infrared Technology，2003，25(3)：13－17.

[3] Liu Dehong，He Lihan，Carin L. Airport detection in large aerial optical imagery[C]// Proceedings of IEEE International Conference on Acoustics，Speech，and Signal Processing，2004(ICASSP'04). Montreal，Quebec，Canada：[s. n.]，2004：761－764.

[4] Gonzalez R C，Woods R E. Digtal Image Processing[M]. Second Edition. Beijing：Publishing House of Electronics Industry，2003：420－440.

[5] YANG Si－hai，CHENG Duan－sheng，XIE Wei－bo. Characteristics of Hough Transform：a Global View[J]. Journal of Computer－aided Design and Computer Graphics，2006，18(8)：1197－1204.

[6] 周得芳，张 健. 二维最大熵阈值分割的一种快速递推算法及应用[J]. 现代电子技术，2003(24)：85－87.

(上接第 189 页)

7890 ABCDEFG"，第三行为"23－68614584 68627404"。打印效果如图 4 所示。

3 结束语

对便携式工业气动标记打印机的硬件组成、移植于 ARM7 上的 μC/OS－II 操作系统进行分析，在此基础上完成了打印机软件方案的确定、软件模块的设计、软件模块编程和软件的调试，完全符合开发便携式打印机的性能要求。

参考文献：

[1] 周立功. ARM 嵌入式系统软件开发实例[M]. 北京：北京航空航天大学出版社，2004.

[2] 胥 静. 嵌入式系统设计与开发实例详解[M]. 北京：北京航空航天大学出版社，2004.

[3] 江卫华. 基于 PC 计算机并行口气动标记系统的设计[J]. 电气传动自动化，2002(1)：44－46.

[4] 李恩林. 插补原理[M]. 北京：机械工业出版社，1984：36－45，112－118.

[5] 林 方. C 语言的汉字处理与图文数据库技术[M]. 西安：西安交通大学出版社，1995.

[6] 黄健青，王 平. Turbo C 矢量字库的分析和使用[J]. 海南大学学报，1995(2)：152－154.

[7] 卢有杰. C 语言常用算法与子程序[M]. 北京：清华大学出版社，1991.

(上接第 192 页)

4 结束语

目前在传统的几种利用人工神经网络进行空气质量预测的方法中，被广泛采用的灰色系统对波动性较强的序列预测效果较差；多层前向网络 BP 算法有收敛速度慢，网络容错能力差，易陷入局部最小的缺点。文中则运用了人工神经网络中的覆盖算法，对从宣城市及周边地区三年的大气环境监测资料、气象资料中得到的数据进行训练学习，建立四季神经计算数据模型，并利用此模型对空气质量进行短期的预测。并将预测结果与实际结果进行比较。实验结果表明：将覆盖算法应用于空气质量预测，效果良好，学习速度快，识别率高，具有较强的实用价值，为实现大气质量预测提供了一种准确高效的方法。

参考文献：

[1] 周志华，曹存根. 神经网络及其应用[M]. 北京：清华大学出版社，2004.

[2] 刘慧婷，倪志伟，李建洋，等. 基于交叉覆盖算法的时间序列匹配[J]. 计算机应用，2007，27(2)：425－427.

[3] 张 铃，张 钹. M－P 神经元模型的几何意义及其应用[J]. 软件学报，1998，9(5)：334－338.

[4] 张 铃，张 钹，殷海风. 多层前向网络的交叉覆盖设计算法[J]. 软件学报，1999，10(7)：737－ 742.

[5] 赵 姝，张燕平，张 媛，等. 基于交叉覆盖算法的入侵检测[J]. 计算机工程与应用，2005(3)：141－143.

[6] 胡光杰，张燕平，陈 洁. 基于覆盖算法的煤炭供应商评测模型[J]. 计算机技术与发展，2007，17(1)：6－8.

[7] 张晨希，张迎春，万 忠，等. 基于交叉覆盖算法的股票预测[J]. 微机发展，2005，15(12)：35－37.