基于BP神经网络的矽肺病预测组合模型研究

章 勤[1,2] 田 晶[1,2] 孙傲冰[1,2] 郑 然[1,2] 陈卫红[3]

(华中科技大学计算机科学与技术学院服务计算技术与系统教育部重点实验室 武汉 430074)[1]

(华中科技大学计算机科学与技术学院集群与网格计算湖北省重点实验室 武汉 430074)[2]

(华中科技大学同济医学院 公共卫生学院环境与健康教育部重点实验室 武汉 430030)[3]

摘 要 矽肺是我国最为严重的职业病之一,严重危害工人的健康。研究表明,矽肺与粉尘接触量、吸烟量、接尘时间等存在明显的剂量反应关系。基于各矽肺致病影响因子,分别利用指数平滑-神经网络ES-BP(Exponential smoothing-BP neural network)和模糊c均值聚类-神经网络FCM-BP(Fuzzy c-means clustering-BP neural network)组合模型对接尘工人未来是否患病以及患病年龄做预测分析。实验结果表明:ES-BP模型能结合原始工人接尘时间队列数据特点,从时间序列上对工人患病年龄进行预测;FCM-BP模型对数据预归类,能极大减小模型复杂度并降低网络训练时间。两种组合模型预测精度均高于BP单独建模预测精度,在工人患病年龄预测中取得了较好的测试效果。

关键词 BP神经网络,指数平滑法,FCM聚类,组合预测,矽肺

中图法分类号 TP183 **文献标识码** A

Research on Hybrid Prediction Methods of Silicosis Based on BP Neural Network

ZHANG Qin[1,2] TIAN Jing[1,2] SUN Ao-bing[1,2] ZHENG Ran[1,2] CHEN Wei-hong[3]

(Services Computing Technology and System Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)[1]

(Cluster and Grid Computing Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China)[2]

(Key Lab of Environment and Health and Department of Occupational and Environmental Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China)[3]

Abstract Silicosis is one of the most harmful occupational respiratory diseases that are threatening the miners working in dust environment in China deadly. Recently the researchers find that pneumoconiosis obeys the actual dose-response relations with the calendar-year exposure matrix, smoking and individual occupational histories. Two hybrid prediction models based on BP neural network, exponential smoothing and FCM clustering were proposed to predict the possibility and the ages of the miners suffering the silicosis. The experiment results show that the efficiency and the accuracy of the both hybrid models are enhanced greatly compared with the single BP neural network; the BP-ES (Exponential Smoothing) model can make a prediction from the time series datum of dust-exposed workers and the other pathogenic factors; and the model complexity and the training time can be highly reduced with the method of pre-classification in BP-FCM clustering model. The hybrid models are effective methods for silica dust diagnosis prediction.

Keywords BP neural network, Exponential smoothing, FCM clustering, Combined forecasting, Cohort of silica dust

1 引言

矽肺是由于工人长期吸入生产性矽尘而引起的以肺组织纤维化为主要特点的疾病,是我国危害最大的职业病之一,至今仍缺乏有效的治疗手段。矽肺与粉尘接触量、吸烟量、接尘时间等存在明显的剂量反应关系[1]。如果能通过有效数据分析方法,结合各矽肺致病影响因子建立矽肺预测模型,对工人未来是否患病、患病时年龄给出科学合理的计算与预测,可以使可能的矽肺患者及早脱离粉尘接触环境,达到预防矽肺发病的目的。

原始矽肺接尘队列追踪时间长,队列中各年接尘值受工人工作时间、作业环境粉尘浓度等复杂非线性因素影响[2],工人患病则与粉尘接触、接尘时间、吸烟等诸多因素密切相关。若基于各致病因子,单独利用BP神经网络模型对工人患病情况做时间序列分析,预测误差会随预测步数的增大而增大;而若仅利用指数平滑、ARMA(Auto-Regressive and Moving

到稿日期:2008-05-05 本文受"863"(项目编号:2006AA02Z347)和"863"(项目编号:2006AA01A115)资助。

章 勤(1955—),女,教授,主要研究方向为图像处理、网格计算,E-mail:qzhang@mail.hust.edu.cn;田 晶(1985—),男,硕士研究生,主要研究方向为数据挖掘、机器学习;孙傲冰(1978—),男,博士研究生,主要研究方向为图像网格、信息集成;郑 然(1977—),女,讲师,CCF会员,主要研究方向为网格计算、网格应用;陈卫红(1966—),女,教授,主要研究方向为职业流行病学。

Average Model)等时序分析方法，则无法揭示各致病因子间的非线性关系。若通过多种预测模型的组合，综合各矽肺患病影响因子，结合各模型自身长处，能更有效地改善模型的拟合能力和提高预测精度。目前大多数组合预测模型都是线性组合，都存在模型权系数选取确定的复杂问题[3]。本文基于BP神经网络模型提出了两种新的组合预测方法：用BP网络确定组合预测模型结构，将指数平滑ES(Exponential smoothing)和模糊C均值聚类FCM(Fuzzy c-means clustering)建模输出分别作为BP网络输入，组合预测模型各输入权重通过网络自学习获得。这样可以避免线性组合预测中各权值计算复杂的问题，又可以充分利用各模型长处，提高预测精度。

本文以中南地区某厂矿中1960年1月1日至1974年12月31日工作一年以上在册职工建立队列，摘录所有队列人员的工作史，综合计算得到每个工人的工作时间-接尘队列[1,4]。从整理得到的工人工作时间-接尘队列中抽取300组样本数据，结合每个工人的加权平均吸烟量、接尘年龄等数据，将指数平滑和聚类分析作为神经网络模型的预处理步骤，分别建立指数平滑-神经网络ES-BP和模糊C均值聚类-神经网络FCM-BP组合预测模型，对工人未来是否患病和患病时年龄做预测。并将组合预测结果与单独利用BP神经网络预测结果进行对照，验证结果显示两种组合建模预测的精度均高于BP单独建模的精度。

2 传统算法描述

2.1 BP神经网络

BP网络基于从输入到输出的映射，对非线性函数的优化进行求解运算。在整个网络学习中，从输入输出数据中提取规律，将其保存于网络权值中并应用于一般情形，具有较强的自学习能力和推广、概括能力。

记η为学习步长，$X_{p1}, X_{p2}, \cdots, X_{pn}$为输入样本，$w_{jn}$为隐含层节点$j$到下一层节点$P_j$的权值，$t_{pk}$为第$k$个输出层的误差纠正因子，则网络结构中隐层节点$j$输出和输入节点$p$的关系为：

$$S_{pj}^h=\sum_i w_{ji}^h x_{pi}, O_{pj}^h=f_j^h(S_{pj}^h)$$

输出节点k和隐层输出节点p的关系为

$$S_{pk}^o=\sum_j w_{kj}^o O_{pj}^h, O_{pk}^o=f_k^o(S_{pk}^o)$$

定义输出误差为δ_{pk}，则$\delta_{pk}=t_{pk}-O_{pk}^o$。

整个网络学习的目的就是为了使如下定义的误差平方和最小：

$$E=\frac{1}{2}\sum_{k=1}^{m}\delta_{pk}^2=\frac{1}{2}\sum_k(t_{pk}-O_{pk}^o)^2$$

在网络训练时，常存在着网络层数选取困难、容易陷入局部极小值等问题。由于BP算法中以解决复杂非线性函数的全局极值为目标，但算法本身是局部搜索的优化方法，训练过程本质上是求非线性函数的极小点问题，这使得它可能陷入局部极值而训练失败。而且随着训练能力的提高，可能出现学习能力下降，引起“过拟合”现象。

2.2 指数平滑法

指数平滑法通过对过去不同时间的资料取不同的权数加权，加以平均以对未来进行预测。利用指数平滑进行时间序列分析时，把距离现在较近的历史数据作为影响较大的因素[5]，同时不断运用误差反馈纠正新的预测值，对短期预测有很好的效果。基于时序$\{X_t\}$的指数平滑公式如下：

$$S_t=\theta X_t+(1-\theta)S_{t-1}, t=1,2,\cdots$$

其中平滑因子$\theta\in[0,1]$；S_t为第t期的指数平滑值。$t>50$时，平滑初始值S_1对S_t计算结果影响极小，取第一期实际值X_1；$t<50$时，初始平滑值S_1对S_t影响较大，取时序$\{X_t\}$前几项的平均值。平滑因子θ决定平滑水平以及对预测值与实际结果间差异的响应速度：θ越接近1，远期实际值对本期平滑值的下降越迅速；θ越接近0，远期实际值对本期平滑值影响程度的下降越缓慢。当时间数列相对平稳时，可取较大的θ；当时间数列波动较大时，应取较小θ，以不忽略远期实际值的影响。

当时间数列无明显的趋势变化时，可用一次指数平滑预测。二次指数平滑适用于具线性趋势的时间数列。三次指数平滑法基于构建抛物线模型，其修正预测值使其跟踪非线性趋势的变化时，广泛用于二次曲线趋势的预测[5]：对于符合$X_t=a+bt+ct^2$的时间序列数据，反复利用指数平滑定义公式可推出a_t, b_t, c_t 3个平滑系数，进而得到3次指数平滑预测值：

$$X_{t+f}=a_t+b_t f+(c_t/2)f^2$$

其中f是预测的时间步长。平滑系数a_t, b_t, c_t值及公式推导过程详见文献[5]。

2.3 FCM模糊聚类

聚类是一个将数据集划分为若干组或类的过程，并使得同一个组内的数据对象具有较高的相似度，而不同组中的数据对象是不相似的。由于同一组中的数据有较高相似度，因此同一个组内的所有对象常常被当作一个对象来进行处理或分析等操作。聚类分析既可以作为一个单独模型对数据分布等特征进行描述，也可以作为其它算法(如分类和定性归纳算法)的预处理步骤。

按照划分结果的不同，聚类可以分为硬聚类和软聚类。软聚类(模糊聚类)中样本按概率可能属于一个或多个聚类结果中，隶属函数或概率是输入样本和聚类中心的关系表述的0～1的值，该输入与所有分类的关系值总和为1。

Bezdek提出的经典模糊C均值聚类(FCM聚类)就是用隶属度确定每个数据点属于某个聚类的程度的一种聚类算法。算法的输出是C个聚类中心点向量和$C*N$的一个模糊划分矩阵，这个矩阵表示的是每个样本点属于每个类的隶属度。根据这个划分矩阵按照模糊集合中的最大隶属原则就能够确定每个样本点归为哪个类。聚类中心表示的是每个类的平均特征，可以认为是这个类的代表点。FCM聚类算法对于满足正态分布的数据聚类效果会很好，另外算法对孤立点是敏感的。算法具体步骤可详见文献[6]。

3 组合预测算法描述

基于传统单一数据建模分析方法对工人患病年龄做预测时，总存在不能充分描述原始工人接尘时间队列数据特性的问题，更不能回避单一模型自身缺陷。因此将多种预测模型通过一定方式组合，可更有效地发挥各模型自身长处，获得优于单一预测模型的预测效果。组合建模预测的数学描述如下：

记$X:\{X_1, X_2, \cdots, X_t, \cdots, X_n\}$为待预测的原始数据，向量$\vec{s_i}$是由模型$i$得到的预测结果($i=1,2,\cdots,k$)，则组合预测

结果为$\vec{s}=\xi(\vec{s_1},\vec{s_2},\cdots,\vec{s_k})$，$\xi()$函数是由组合方法确定的组合预测函数，可以是 k 个模型分别预测后对所得预测结果的加权组合，也可把模型 i 的输出作为模型 $i+1$ 的输入进行模型迭加。本文所用 2 种组合预测建模方法均属于后者。

3.1 ES-BP 组合预测模型

ES-BP 组合预测基本思想是先利用指数平滑法对接尘时间队列数据进行预测得到工人年接尘预测值，再把得到的接尘预测值与接尘年龄和加权平均吸烟量等致病因子一起输入 BP 网络进行训练，预测工人是否患病以及患病时年龄。这样既可以利用指数平滑法充分结合原始矽肺接尘时间队列数据特点，又可以通过神经网络描述各矽肺影响因子间的非线性结构关系，还可以弥补单独利用 BP 做时间序列预测时由于时序期数太长带来的误差累积问题，实现 2 个模型的优势互补，提高预测精度。ES-BP 组合预测模型结构如图 1 所示。

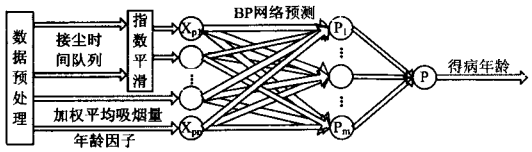

图 1　指数平滑-BP 组合模型结构

具体的组合建模预测步骤如下：

1)指数平滑预测接尘值。选取工人接尘时间队列数据，确定合适的指数平滑因子 a 及合适的指数平滑次数后对工人未来接尘值做预测。

2)BP 网络预测患病年龄。把由指数平滑得到的接尘预测值和吸烟队列数据、年龄因子一起经过归一化后作为网络输入，工人是否患病以及患病时年龄作为网络输出，对训练数据做网络训练。

3)应用训练好的网络对预测样本数据做预测分析。网络输出为 0 时代表工人不患病，为其它数值时代表工人患病年龄的预测值。

3.2 FCM-BP 组合预测模型

基于聚类的 FCM-BP 组合预测模型的基本思想是先利用聚类算法把训练样本划分成几类，使各类样本保持较多的相似性，再用神经网络对各类样本分别训练，并将训练得到的网络模型分别对归类后的预测样本做判断。这样可以有效去除原始数据中的异常点，提高神经网络训练模型的训练精确度并减小模型的复杂度[7]。

由于在做神经网络预测之前先对训练样本进行了聚类分析，把大的数据集划分成为相似的较小的几类，同时在每一小类中，数据保持较高的相似度，这样在对每一小类进行网络训练时肯定会降低训练的次数，网络模型的复杂度会降低很多，网络的收敛速度会更快。在之后进行 BP 网络训练时，就可以减少中间层最佳中心向量的个数，使原来较复杂的模型转换为几个不同的较简单的模型。经过聚类分析后，还可以使原始数据中的异常点被分为数据集较少的一类或几类，不参加神经网络的训练，使神经网络能更好地近似原始时间队列数据的规律。

FCM-BP 组合预测模型结构如图 2 所示。

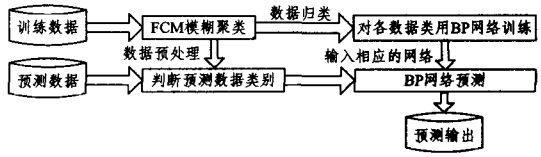

图 2　FCM-BP 组合模型结构

FCM-BP 组合预测模型步骤如下：

1)对训练样本和预测样本分别做数据预处理。由于原始接尘时间队列数据随时间推移有明显递减趋势，在保证矽肺数据一致性的基础上，对输入的接尘时间队列数据去除明显偏离趋势曲线的采样点后，结合吸烟、接尘年龄等致病因子分别归一化到[0.1,0.9]区间。

2)通过 FCM 聚类算法对训练样本数据进行聚类划分，得到 N 个小类$\{C_1,C_2,\cdots,C_n\}$，利用 BP 网络分别对这 N 个类$\{C_1,C_2,\cdots,C_n\}$中的训练样本数据做网络训练，得到 N 个训练好的 BP 网络$\{P_1,P_2,\cdots,P_n\}$。

3)对预测样本数据应用 FCM 聚类算法判定所属类 C_i。分别应用类 C_i 中训练得到的 BP 网络 P_i 对该类中的预测样本数据做预测，得到预测结果。

4　建模预测分析

本文对 ES-BP 和 FCM-BP 组合预测模型分别进行编程实现。实验数据选用同一厂矿中 300 组工人的矽肺队列数据，包含时间序列上的多种矽肺影响因子，包括粉尘接触量、吸烟量、接尘年龄(由接尘时间平移得到)等等，对工人未来是否患病和患病时年龄进行组合建模预测分析。

4.1 ES-BP 组合建模预测

对原始时间队列中 300 组工人接尘时间队列数据利用指数平滑方法预测未来接尘值，得到每个工人的接尘预测值。矽肺接尘时间队列样本中年接尘值随时间变化具有非线性递减趋势，采用 3 次指数平滑法预测工人未来接尘值。实验比较知平滑因子 a 取 0.6 较合适。由于接尘年数小于 50，故第一年预测值取前三年实际值的平均。

将指数平滑预测的工人接尘值和加权平均吸烟量、接尘年龄归一化后作为 BP 网络输入。BP 网络隐含层数和初始权值通过网络训练自学习获得，学习速率取 0.1。激活函数为双极型激活函数 Sigmoid 函数，期望误差设为 0.001。其中训练数据取 250 组，预测数据取其余 50 组。输出为 0 时表示工人不患病，为其它数值时表示工人患病时年龄预测值。将工人患病年龄预测值和实际患病年龄比照得到个体预测误差，对 50 组预测样本中的个体预测误差求平均得到组合建模预测误差[8]。

4.2 FCM-BP 组合预测模型

FCM-BP 组合模型在做预测前，需要先对原始接尘时间队列数据做必要预处理，将 300 组工人的接尘时间队列数据中各年接尘数据值求平均，得到 300 组工人个体的年平均接尘值数据。将求得的年平均接尘值和加权平均吸烟量、接尘年龄等致病因子一起作为 FCM-BP 组合预测模型的输入，把工人患病时年龄作为输出，通过 BP 网络训练对工人患病年龄做预测。其中训练数据取 200 组，预测数据取其余 100 组。利用训练得到的分类结果对预测数据进行分类判断后分别应

用不同的网络结构做预测。

组合模型中聚类数设定为 5，BP 网络训练参数与 ES-BP 中各训练参数相同。网络输出为 0 时表示工人不患病，为其它数值时表示工人患病年龄的预测值。

4.3 组合模型建模与单独 BP 建模分析比较

对 300 组工人矽肺队列数据分别用 2 种组合模型对工人患病年龄做预测，从预测精度和模型复杂度角度分别对两种组合模型做评估，将组合建模预测与单独使用 BP 建模预测做比较。

图 3 给出采用 ES-BP 组合模型与单独使用 BP 网络对 50 组预测样本数据中工人患病年龄的预测情况。由图中看出 ES-BP 组合模型预测值比 BP 预测更接近工人实际患病年龄。

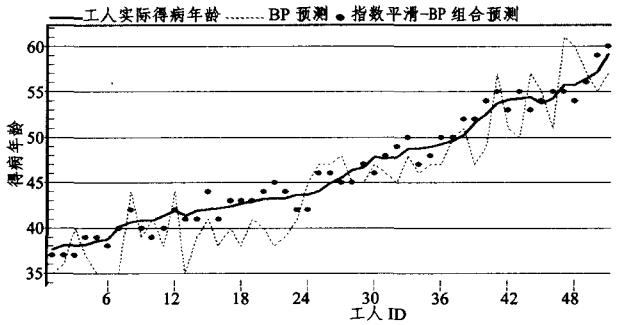

图 3 ES-BP 预测与 BP 预测

表 1 将 FCM-BP 组合建模与 BP 单独建模训练开销做比较。由表中数据分析得知，由于 FCM-BP 组合模型做预测时先对数据进行类别判断，使每一小类数据集具有较大相似性，因此对每小类做网络训练时数据集数目相对减少，BP 网络训练次数、运行时间和模型复杂度都减小很多。

表 1 FCM-BP 组合建模执行效率分析

预测模型	BP	FCM 聚类-BP 组合	
		类 1	350
		类 2	400
训练次数	5000	类 3	370
		类 4	450
		类 5	475
运行时间/s	75.35	45.26	

表 2 显示了模型评价的结果。分别计算组合建模与 BP 单独建模预测的平均相对误差(MRE)、均方误差(MSE)、平均绝对误差(MAE)和残差平方和(SSE)，表明两种组合建模预测的误差都低于 BP 模型预测的误差。

表 2 组合建模预测与 BP 预测误差分析

患病年龄预测	BP	指数平滑-BP	FCM 聚类-BP
MRE	16.1732%	12.3684%	13.4763%
MSE	0.1576	0.1274	0.1463
MAE	0.1453	0.1358	0.1164
SSE	0.1475	0.1263	0.1183

结束语 本文基于粉尘接触量、吸烟量、接尘时间等影响矽肺的各致病因子，研究了分别采用 ES-BP 和 FCM-BP 组合模型的矽肺预测方法，并和单一采用 BP 建模预测进行了对比。经实验比较分析知：ES-BP 组合模型能结合矽肺时间队列数据的非线性特性，综合各种致病因素，从时间段上对工人患病年龄做出比 BP 更准确的预测；FCM-BP 组合模型对矽肺数据进行聚类，显著降低了网络训练的复杂度和训练时间。

由此可知，影响因子较多且非线性联系显著的时间序列数据，更适合应用组合模型进行预测，且两种组合模型预测精度均高于单独使用 BP 建模的预测精度，有较好的应用价值。

参 考 文 献

[1] 陈卫红，张小康，王海椒，等. 锡矿作业工人粉尘接触和队列死因分析[J]. 环境与职业医学，2007(24)：9-12

[2] Scalia A P C, Barreto S M, Siqueira A L, et al. Continued Exposure to Silica After Diagnosis of Silicosis in Brazilian Gold Miners [J]. American Journal of Industrial Medicine, 2006, 49: 811-818

[3] 甘健胜，陈国龙. 线性组合预测模型及其应用[J]. 计算机科学，2006，33(9)：191-194

[4] Chen W, Eva H, Chen J Q. Risk of Silicosis in Cohorts of Chinese Tin and Tungsten Miners, and Pottery Workers (I): An Epidemiological Study [J]. American Journal of Industrial Medicine, 2005, 48: 1-9

[5] 何书元. 应用时间序列分析[M]. 北京：北京大学出版社，2004

[6] 高新波. 模糊聚类分析及其应用[M]. 西安：西安电子科技大学出版社，2004

[7] 邓赵红，王士同. 鲁棒性的模糊聚类神经网络[J]. 软件学报，2005，16(8)：1415-1422

[8] 杨奎河，王宝树，赵玲玲. 基于神经网络的预测模型中输入变量的选择[J]. 计算机科学，2003，30(8)：139-140，143

(上接第 223 页)

[2] Wang R. A Product Perspective on Total Data Quality Management[J]. Communications of the ACM, 1998, 41(2): 58-65

[3] Naumann F. From Databases to Information Systems-Information Quality Makes the Difference // Proc. of the International Conference on Information Quality (IQ). USA, 2001

[4] Barnes S, Vidgen R. Assessing the Quality of Auction Web Sites//Proc. of the 34th Annual Hawaii International Conference on System Sciences. USA, 2001

[5] 洪月华. 基于模糊综合评价的课堂教学质量数据挖掘[J]. 计算机科学，2008，35(2)：154-156，170

[6] Olsina L, Rossi G. Measuring Web Application Quality with WebQEM[J]. IEEE MultiMedia, 2002, 9(4): 20-29

[7] Mich L, Franch M, Gaio L. Evaluating and Designing Web Site Quality[J]. IEEE MultiMedia, 2003, 10(1): 34-43

[8] 朱焱，唐慧佳，马永强. 基于 ISO/IEC9126 的 Web 资源质量评测系统[J]. 西南交通大学学报，2008，43(2)：253-257

[9] Zhu Y. Group Assessment of Web Source/Information Quality Based on WebQM and Fuzzy Logic//G. Wang, et al., eds. Lecture Notes in Artificial Intelligence, vol. 5009. Germany: Springer Verlag, 2008: 660

[10] Saaty T. The Analytic Hierarchy Process. McGraw-Hill, Inc., 1980