一种改进的径向基神经网络预测算法①

王行甫，覃启贤，程用远，侯成龙

(中国科学技术大学 计算机学院, 合肥 230027)

摘 要：神经网络是数据挖掘的常用的方法之一，主成分分析方法是统计学多元分析中的一种分析多个变量间内在关系的方法。将主成分分析预处理方法与神经网络结合起来使用，可以分析原始变量间关系，将原始数据降维，减少数据规模。对神经网络算法和主成分分析相关理论进行了研究，在此基础上，结合大量的气象数据和北京的传染病数据，提出了一种改进的基于主成分分析预处理结合神经网络算法的数据挖掘方法。通过对比实验测试，本文提出的组合算法在收敛速度及预测准确性方面的性能有了很大程度提高。结合国家重大专项疾病预测项目，将该方法应用于其中的流行性传染病的预测上。

关键词：数据挖掘；径向基神经网络；主成分分析

Improved RBF Neural Network Prediction Algorithm

WANG Xing-Fu, QIN Qi-Xian, CHENG Yong-Yuan, HOU Cheng-Long

(School of Computer Science, University of Science & Technology of China, Hefei 230027, China)

Abstract：The neural network is a kind of the commonly used method of data mining, principal component analysis method is a kind of method that analyzes internal relationship between the many variables of the multivariate analysis of statistical. Combined the principal component analysis pretreatment method with neural network, you can analyze the relationship between the original variables, reduce dimensions of the original data and reduce the scale of data. This paper does research on the neural network algorithm and the principal component analysis correlative theory. Based on this, combined with a large number of meteorological data and disease data of Beijing, we proposed an improved method of the data mining which based on principal component analysis and neural network algorithm preprocessing. Through the contrast experiment test, the combinations of the algorithm have a large degree increase in the convergence rate and forecast accuracy property.

Key words：data mining; RBF neural network; principal component analysis

1 概述

进入 21 世纪以来，世界各地非典、甲流等群体性传染病大量出现，我国中医学专家根据古代中医理论成功预测了疫情的爆发消亡时间，并指出疫情的爆发和全国的气候出现的异常存在很大的关联性，为了增强对疫情的预测能力，我们有必要用科学的建模方法分析气象因素对疾病的影响，进一步解析气候与疾病的非线性关系。影响因素有很多，包括气温、气压、湿度、降水量、日照、风速等等多种变量。只有从中找出一组合适的输入才能有效地解释疾病的发病人数的变化，才能对疾病的发病趋势做出准确预测[1]。

神经网络是数据挖掘特别是在预测方面常用的方法之一，可以较好地优化网络结构，但是没有考虑输入变量的选取。输入变量过多时，网络结构复杂，加重了神经网络的训练负担，学习速度急剧下降；同时，主观选择很有可能包含与输出相关性很小的输入变量，增加了陷入局部极小点的可能性，不但不能提高预测精度，反而降低了神经网络预测的性能[2]。

① 基金项目:国家科技重大专项(2012ZX10004-301-609);国家自然科学基金(60970128)
收稿时间:2011-12-02;收到修改稿时间:2012-03-06

针对上述现象，本文提出了基于主成分分析—RBF 神经网络的疾病预测模型。该模型首先利用主成分分析技术将影响疾病预测的众多因素变量进行分析变换，有效消除原训练样本空间的信息重叠和噪声，尽可能多地保留原有数据的有用信息，降低数据维度，减小网络规模，得到一组彼此不相关的新输入变量，然后将重构的训练样本空间作为 RBF 神经网络的输入，进行疾病预测。

2 主成分分析方法

主成分分析[3,4]方法的主要思想是将多个变量通过线性变换以选出少数几个重要变量的一种多元统计分析方法，是因子分析方法的一种，它将原来众多具有一定相关性的指标重新组合成一组新的相互无关的综合指标来代替原来指标的统计方法，也是数学上处理降维的一种方法。

2.1 对样本数据的标准化

设有 n 个样品，p 个指标的样本原始数据。

$$Y=\begin{bmatrix} y_{11}y_{12}\cdots y_{1p} \\ y_{21}y_{22}\cdots y_{2p} \\ \cdots\cdots\cdots \\ y_{n1}y_{2n}\cdots y_{np} \end{bmatrix}$$

对数据矩阵 Y 进行标准化处理，即对每一个指标分量作标准化变化，变换公式为：

$$X_{ij}=\frac{Y_{ij}-\overline{Y}_j}{S_j} \qquad \begin{pmatrix} i=1,2,\cdots,n \\ j=1,2,\cdots p \end{pmatrix}$$

其中，样本均值 $\overline{Y}=\frac{1}{n}\sum_{k=1}^{n}Y_{ki}$，样本标准差 $S_i=\sqrt{\frac{1}{n=1}\sum_{k-1}^{n}(Y_{ki}-\overline{Y}_i)^2}$

得到标准化后的数据矩阵 $X=\begin{bmatrix} x_{11}x_{12}\cdots x_{1p} \\ x_{21}x_{22}\cdots x_{2p} \\ \cdots\cdots\cdots \\ x_{n1}x_{n2}\cdots x_{np} \end{bmatrix}$

2.2 计算相关矩阵

对于给定的 n 个样本，求样本间的相关系数。相关矩阵中的每一个元素由相应的相关系数所表示。

$$R=XX'=\begin{pmatrix} 1r_{12}\cdots r_{1p} \\ r_{21}1\cdots r_{2p} \\ \cdots\cdots\cdots \\ r_{p1}r_{p2}\cdots 1 \end{pmatrix}$$

其中 $r_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}X_{ki}X_{kj}$

2.3 求特征值和特征向量 λ

假设求得的相关矩阵为 R，求解特征方程：

$$/R-\lambda i/=0$$

通过求解特征方程，可得到 m 个特征值($i=1\sim m$)，和对应于每一个特征值的特征向量：

$$a_i=(a_{i1},a_{i2},\cdots,a_{imp}) \quad i=1\sim m$$

且有 $\lambda_1\geq\lambda_2\geq\lambda_3\geq\lambda_m\geq 0$

假设相应 λ_1 的特征向量 $A_i=(a_{1i},a_{2i,\cdots,}a_{pi})$

2.4 求主成分

根据求得的 m 个特征向量，m 个主成分分别为：

$$\begin{aligned} F_1&=a_{11}X_1+a_{12}X_2+\cdots+a_{1p}X_p \\ F_2&=a_{21}X_1+a_{22}X_2+\cdots+a_{2p}X_p \\ &\cdots\ \cdots\cdots\ \cdots\cdots\ \cdots\cdots\cdots \\ F_m&=a_{m1}X_1+a_{m2}X_2+\cdots+a_{mp}X_p \end{aligned}$$

求各主成份的关键是求特征根及其相应的特征向量。

主成份分析以较少的 m 个指标代替了原来的 p 个指标对系统进行分析，这给我们的综合评价带来了很大的方便。

保留多少主成分取决于保留部分的累积方差在方差总和中所占百分比(即累积贡献率)，规定一个百分比便可以决定保留几个主成分，当多留一个主成分，累积方差变化不大的时候，便不再多留。

在本文中取累积贡献率达到 85%以上时的因子个数。

3 改进的径向基函数神经网络预测算法

1988 年,Moody 等首次提出径向基函数(Radial Basis Function，RBF) 神经网络[5,6] ，该网络的优点在于能以任意精度逼近任意连续函数，因而在复杂非线性输入-输出系统建模中得到了广泛应用。RBF 神经网络通常具有三层网络结构，分别为输入层、隐含层、输出层。输入层神经元将输入信号传递到隐含层，隐含层神经元通过基函数对输入信号产生局部响应，当输入信号靠近基函数的中心时，隐含层神经元将产生较大的输出，反之，则产生较小的输出，输出层神经元通常采用简单的线性函数产生网络输出信号。

RBF 网络常用的基函数为高斯基函数，激活函数可以表示为：

$$R_i(x_p - c_i) = exp\left(-\frac{1}{2\sigma^2}\left\|x_p - c_i\right\|^2\right)$$

式中$\left\|x_p - c_i\right\|$为欧氏范数，c为高斯函数的中心，σ为高斯函数的方差。由图 1 的径向基神经网络的结构可得到网络的输出为：

$$y_i = \sum_{i=1}^{h} w_{ij} exp\left(-\frac{1}{2\sigma^2}\left\|x_p - c_i\right\|^2\right)\ \ j = 1,\ \cdots,\ n$$

式中$x_p = (x_1^p, x_2^p, \cdots x_3^p)^T$为第 p 个输入样本，c_i为网络隐含层结点的中心，w_{ij}为隐含层到输出层的连接权值，i = 1,2，…，h为隐含层的节点数，y_i为与输入样本对应的网络的第 j 个输出结点的实际输出。设 d 是样本的期望输出值，那么基函数的方差可表示为：

$$\sigma_i = \frac{1}{P}\sum_{j}^{m}\left\|d_j - y_j c_i\right\|^2$$

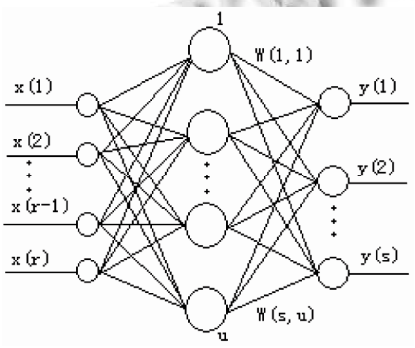

图 1　径向基神经网络结构

按照基函数中心选取方法不同可以将 RBF 网络学习方法分为随机算法、自组织学习法、最近邻聚类学习法及正交最小二乘法等[7]。根据本研究的特点，以最近邻聚类算法了基础，设计了一种自适应基函数选择算法。

算法具体步骤如下：

(1) 改进的基函数中心选取方法

a) 根据数据的特点，选择一个合适的参数r.从第一个数据$(x1,y1)$开始$c1=(x1,y1)$。这时的 RBF 网络，只有一个隐含层单元，中心为 c1。

b) 考虑第 2 个数据$(x2,y2)$，求出第 2 个数据到第 1 个数据的欧式距离L，如果$L <= r$，则$c1$为第 2 个数据的中心；如果$L > r$，则第 2 个数据成为一个新的聚类中心$c2$。

c) 考虑第 k 个数据(xk,yk)，此时已经有了 N 个聚类中心，分别为：$c1$，$c2$,…,cN，。分别求出第 k 个数据和这 N 个聚类中心的欧氏距离，设 L 为这些距离中的最小值。如果$L <= r$，则保持聚类中心个数不变；如果$L > r$，则第 k 个数据成为新的聚类中心。

r 是一个一维参数，可以通过反复的实验来选择一个合适的值。

(2) 求解方差

我们选取的 RBF 神经网络的基函数为高斯函数，因此求得方差σ_i如下所示：

$$\sigma_i = \frac{c_{max}}{\sqrt{2h}}\ \ i = 1,2,\cdots,h$$

其中，c_{max}表示所选取中心之间的最大距离。

(3) 计算隐含层和输出层之间的权值。

隐含层和输出层之间神经元的连接权值可以用最小二乘法直接计算得到，计算公式如下：

$$w = exp\left(\frac{h}{c_{max}^2}\left\|x_p - c_i\right\|^2\right)\ \ p = 1,2,\cdots,P; i = 1,2,\cdots,h$$

4　预测结果分析

用主成分分析方法对影响月发病人数的气象因素：月平均气温(X_1)，月平均气压(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)，月平均风速(X_6)进行分析，分析前先进行数据标准化处理。对北京 2004-2009 年的气象数据进行主成分分析得到表 1 和表 2，由表 1 中数据可以看出前 2 个因子的贡献率就达到了 87.18%(一般取贡献率大于 80%的成份)，因此取 m = 2 作为主成分神经网络模型输入因子的维数。

表 1　特征根与方差贡献表

成份	初始特征值			提取平方和载入		
	合计	方差%	累积%	合计	方差%	累积%
1	3.30	54.99	54.99	3.300	54.99	54.99
2	1.93	32.18	87.18	1.931	32.18	87.18
3	.435	7.257	94.43			
4	.232	3.873	98.31			
5	.078	1.295	99.60			
6	.024	.393	100.0			

表 2 为输入因子的特征向量矩阵，由表 2 我们可以得到输入向量的表达式为：

Y1 = 0.867X1-0.804X2+0.901X3+
0.871X4-0.303X5-0.488X6
Y2 = 0.423X1-0.552X2-0.373X3+
0.167X4+0.857X5+0.739X6

表 2 特征向量矩阵

	成份	
	1	2
X1	0.867	0.423
X2	-0.804	-0.552
X3	0.901	-0.373
X4	0.871	0.167
X5	-0.303	0.857
X6	-0.488	0.739

取 2004 年至 2008 年的数据作为预测模型的训练样本，用 matlab9.0 建立 RBF 神经网络模型对北京的肺结核病的 2009 年前五个月发病人数做预测，经过不断的调整参数和反复的实验，预测的结果如表 3 所示，此外还用传统 RBF 神经网络做了对比。

表 3 预测结果

	1	2	3	4	5
实际值	0.6633	0.5899	0.7188	0.6867	0.6513
改进 RBF	0.6520	0.5466	0.7144	0.7288	0.6650
传统 RBF	0.6997	0.5145	0.7389	0.7608	0.7882

从上面的发病人数的实际值和预测值的对比结果可以看出，用训练好的 RBF 网络对北京得肺结核发病人数可以做出很好的预测，虽然个别预测值与实际值有差别，但是完全可以体现出发病的整体趋势，可以在发病趋势增大时提供预警，做好相关的预防准备。

5 结论

径向基神经网络属于前向神经网络类型，它以局部响应的径向基函数代替传统的全局响应的激励函数，免了传统 BP 学习算法的收敛时间长，易陷入局部最优的缺点。径向基神经网络预测系统具有良好的性能，我们将数理统计中的主成分分析方法结合到预测模型中来，进一步提高了预测的精度。仿真结果显示，主成分-RBF 神经网络模型具有很好的预测精度，可以满足实际需求。

参考文献

1 Tian FP, Wan SH. Principal component neural network prediction model for disease prediction.Journal of Liaoning Technical University(Natural Science),2010,(5).

2 Zhi J, Zhang DM. Based on PCA of genetic neural network prediction of stock index. Computer Engineering and Applications,2009,45(26):210−212.

3 Xiang DJ. Practical Multiply Statistical Analysis. China University of Geosciences Press,2005.120−184.

4 Cben TP, Lin Q. New algorithm for pincipal component and minor component extraction.Journal of Fudan university (Natural Science),1997,36(2):227−230.

5 Moody JE, Darken C. Fast learning in networks of locally tuned processing units. Neural Computation,1989,1(2):281−294.

6 Andrea B, Evan B, Thomas MDA, et al. An overview and sensitivity study of a multi-mechanistic chloride transport model. Cement Concrete Res,1999,29(6):827.

7 Robert J, Schilling James, Carroll J. Approximation of non-linear systems with radial basis function neural networks. IEEE Trans. on Neural Networks,2001,12(1):21−28.

(上接第 109 页)

法.生物识别研究新进展(二).北京:清华大学出版社,2003. 192−196.

3 杨钦.限定 Delannay 三角网格剖分技术.北京:电子工业出版社,2005.

4 刘宁,邵晓艳,李向.DT 网格在指纹识别中的应用.河南机电高等专科学校学报,2008,16(4):34−36.

5 Qi J, Wang YS. A robust fingerprint matching method. Pattern Recognition,2005,38:1665−1671.

6 张季.自动指纹识别算法研究与系统设计[硕士学位论文].成都:西南交通大学,2007.