基于 PCA 和改进 BP 网络的降雨预报模型研究

刘 乐[1],王洪国[1,2],王宝伟[3]
LIU Le[1],WANG Hong-guo[1,2],WANG Bao-wei[3]

1.山东师范大学 管理与经济学院,济南 250014
2.山东省科学技术厅,济南 250011
3.山东师范大学 信息科学与工程学院,济南 250014
1.School of Management and Economy, Shandong Normal University,Ji'nan 250014,China
2.Department of Science and Technology of Shandong Province,Ji'nan 250011,China
3.School of Information Science and Engineering, Shandong Normal University,Ji'nan 250014,China
E-mail:lele_double@163.com

LIU Le,WANG Hong-guo,WANG Bao-wei.Research in rain forecasting model based on PCA and improved BP network.Computer Engineering and Applications,2008,44(12):234-237.

Abstract: On the base of combining Principal Component Analysis with improved BP network,this paper made a research on the rain forecasting model.First the dimensions of the raw meteorological data were decreased by PCA.Then it was by improved BP network to learn the potential rules which existd in meteorological samples effectively.The result of the research shows that, the rain forecasting model has high training efficiency and good forecasting effect.

Key words: Principal Component Analysis(PCA);Back-Propagation network;rain forecasting

摘 要:在主成分分析法和改进 BP 网络相结合的基础上,进行降雨预报模型的研究。先由主成分分析法降低原始气象数据的维数,然后利用改进 BP 网络有效地学习气象样本数据中蕴含的内在规律。研究结果显示,该降雨预报模型训练效率高,预报效果好。

关键词:主成分分析;BP 网络;降雨预报
文章编号:1002-8331(2008)12-0234-04 **文献标识码:**A **中图分类号:**TP183

1 引言

在现代社会中,天气预报对于工农业生产、交通安全、室外作业质量、防灾救灾等有着非常重要的作用,会直接关系到经济效益、工程进度、产值收成、甚至生命财产安全。降雨预报作为天气预报的重点和难点,受气压、气温、湿度、风向和风速等多方面因素的影响。这些因素之间存在较强的相关性,直接纳入分析不仅复杂,而且因素间难以取舍。主成分分析法能够有效地解决这个问题,通过线性变换将原本多个指标组合成相对独立的、能充分反映总体信息的少数几个指标,以利于接下来的研究。

自从 1987 年, 美国的 Neural Ware 公司开发出第一个基于人工神经网络的天气预报系统并取得较好效果以来,神经网络(尤其是 BP 网络)凭借其联想记忆、并行处理、自学习、自组织等特点迅速成为天气预报领域的重要工具和研究热点。文献[2]中采用离散值的 BP 网络进行降雨预报, 但预报准确率仅为 50%左右。文献[3]中采用增加动量项的 BP 网络来预报降雨并取得不错的效果,然而加动量项的 BP 网络依然还有很大的性能提高空间。本文的工作就是将主成分分析法与采用连续值的新型改进 BP 网络有机地结合起来, 应用于降雨预报的研究中,并揭示出这种新型降雨预报模型的可行性和实用价值。

2 主成分分析法的基本原理

主成分分析法(Principal Components Analysis)是一种通过降维来简化数据结构的方法,主要用于多变量问题中。提取出来的每个主成分是原来多个变量的线性组合。用数学语言可以简述如下:

设 $\boldsymbol{X}^{\mathrm{T}}=(X_1,X_2,\cdots,X_p)$为 P 维随机向量,每个随机变量 X_i $(i=1,\cdots,p)$有 n 个样本,它的主成分为:

$$\begin{aligned}\boldsymbol{F}_1&=\boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{X}=e_{11}X_1+e_{21}X_2+\cdots+e_{p1}X_p\\&\vdots\\\boldsymbol{F}_p&=\boldsymbol{e}_p^{\mathrm{T}}\boldsymbol{X}=e_{1p}X_1+e_{2p}X_2+\cdots+e_{pp}X_p\end{aligned}\tag{1}$$

主成分的方差决定着它反映总体信息的多少, 第一主成分 $\boldsymbol{F}_1$ 在所有线性组合中方差最大,后面依次是 $\boldsymbol{F}_2,\boldsymbol{F}_3,\cdots,\boldsymbol{F}_p$。值得注意的是,各主成分之间的相关系数为 0。

进行主成分分析之前,通常需要先计算原始随机变量的相

基金项目:山东省自然科学基金(the Natural Science Foundation of Shandong Province of China under Grant No.Q2006G03)。
作者简介:刘乐(1984-),男,硕士研究生,主要研究领域为数据挖掘,神经网络;王洪国(1966-),男,博士后,教授,主要研究领域为组合优化算法,数据挖掘,电子政务;王宝伟(1981-),男,硕士研究生,主要研究领域为遗传算法,人工神经网络。
收稿日期:2007-08-29 **修回日期:**2007-10-22

关矩阵。设样本数据矩阵为 $\boldsymbol{M}=(\alpha_1,\alpha_2,\cdots,\alpha_p)^{\mathrm{T}}=(\beta_1,\beta_2,\cdots,\beta_n)$，其中 $\beta_j(j=1,\cdots,n)$对应于每个随机向量 $\boldsymbol{X}$ 的样本向量，$\alpha_i(i=1,\cdots,p)$对应于每个随机变量的所有样本值构成的向量，则它的样本方差-协方差矩阵 $\boldsymbol{S}$ 和相关矩阵 $\boldsymbol{R}$ 分别为：

$$\boldsymbol{S}=(S_{ab})=\frac{1}{n}\sum_{k=1}^{n}\left[(M_{ak}-\frac{1}{n}\sum_{c=1}^{n}M_{ac})(M_{bk}-\frac{1}{n}\sum_{d=1}^{n}M_{bd})\right] \tag{2}$$

$$\boldsymbol{R}=(R_{ab})=\frac{S_{ab}}{\sqrt{S_{aa}}\sqrt{S_{bb}}} \tag{3}$$

为使主成分分析能均等地对待每一个原始随机变量，消除由于单位的不同而可能带来的不利影响，须将各原始随机变量作标准化处理，得到 $X_i^*(i=1,\cdots,p)$。$\boldsymbol{X}^*$ 的样本方差-协方差矩阵 $\boldsymbol{S}$ 即为相关矩阵 $\boldsymbol{R}$。

由相关矩阵 $\boldsymbol{R}$ 求出它的特征值 $\lambda_i(i=1,\cdots,p)$，根据特征值的大小可确定主成分的次序并计算出各主成分的方差贡献率。此时，按顺序取前 $m(m\leqslant p)$个主成分，使其累计贡献率达到一个较高的百分数。接着逐一计算出选定主成分 $\boldsymbol{F}_t(1\leqslant t\leqslant m)$和标准化随机变量 X_i^* 之间的相关系数(即因子负荷量)$\rho(Y_t,X_i^*)$，并由

$$\rho(Y_t,X_i^*)=\sqrt{\lambda_i}\cdot e_{ti} \tag{4}$$

计算出 e_{ti}，从而得到 $\boldsymbol{F}_t$ 与原始随机向量 $\boldsymbol{X}$ 中 P 个变量的线性组合关系。事实上，主成分 $\boldsymbol{F}_t$ 的系数向量为第 t 个特征值 λ_t 所对应的正交化特征向量 $\boldsymbol{e}_t$。

3　BP 网络的基本原理

BP(Back Propagation)网络[1]是按照误差梯度下降原则进行权值调整的多级前馈神经网络，由输入层、隐层、输出层组成，具有很强的非线性映射能力、泛化能力和容错能力，能较好地处理非线性决策问题。

在 BP 网络中，单隐层 BP 网络应用最为广泛。如图 1 所示，其中输入向量为 $\boldsymbol{X}=(x_1,x_2,\cdots,x_i,\cdots,x_n)^{\mathrm{T}}$，隐层输出向量为 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_j,\cdots,y_m)^{\mathrm{T}}$，$x_0=-1$ 和 $y_0=-1$ 是为隐层节点、输出层节点引入阈值而设置的，$\boldsymbol{O}=(o_1,o_2,\cdots,o_k,\cdots,o_l)^{\mathrm{T}}$ 表示输出层输出向量，$\boldsymbol{d}=(d_1,d_2,\cdots,d_k,\cdots,d_l)^{\mathrm{T}}$ 为期望输出向量，$\boldsymbol{V}=(V_1,V_2,\cdots,V_j,\cdots,V_m)$和 $\boldsymbol{W}=(W_1,W_2,\cdots,W_k,\cdots,W_l)$分别表示输入层和隐层之间、隐层和输出层之间的权值矩阵。

各层信号之间存在以下数学关系：对于输出层，有

$$o_k=f(net_k)\quad net_k=\sum_{j=0}^{m}w_{jk}y_j\quad k=1,2,\cdots,l \tag{5}$$

对于隐层，有

$$y_j=f(net_j)\quad net_j=\sum_{i=0}^{n}v_{ij}x_i\quad j=1,2,\cdots,m \tag{6}$$

BP 网络所采用的 BP 算法包括信号正向传播、误差反向传播两个过程。一般取

$$E=\frac{1}{2}\sum_{q=1}^{P}\sum_{k=1}^{l}(d_k^q-o_k^q)^2 \tag{7}$$

作为网络误差度量，其中 P 代表训练样本的总数，q 代表第 q 个训练样本。根据图 1 所示，标准 BP 算法的过程可简述如下：

(1)初始化：对所有权值赋以随机小数，并对阈值设定初值；

(2)给定训练样本集：提供输入向量 $\boldsymbol{X}$ 和期望输出向量 $\boldsymbol{O}$；

(3)计算实际输出 $\boldsymbol{Y}$：$y_j=f(net_j)$，其中 $f(\cdot)$为 Sigmoid 函数，即 $f(x)=\dfrac{1}{1+e^{-x}}$；

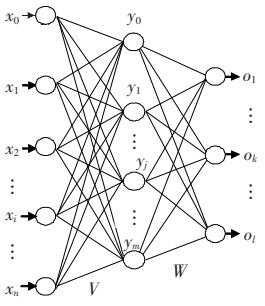

图 1　单隐层 BP 网络

(4)调整权值：按误差反向传播的方向，从输出层到输入层按下式修正权值：

$$\begin{cases}\Delta w_{jk}=\eta\cdot(d_k-o_k)\cdot o_k\cdot(1-o_k)\cdot y_j\\ \Delta w_{ij}=\eta\cdot\sum_{k=1}^{l}(w_{jk}\cdot\delta_k)\cdot o_j\cdot(1-o_j)\cdot x_i\end{cases} \tag{8}$$

若加入动量项，权值按式(9)进行调整；

$$\begin{cases}w_{jk}(t+1)=w_{jk}(t)+\Delta w_{jk}+\alpha[w_{jk}(t)-w_{jk}(t-1)]\\ w_{ij}(t+1)=w_{ij}(t)+\Delta w_{ij}+\alpha[w_{ij}(t)-w_{ij}(t-1)]\end{cases} \tag{9}$$

(5)返回(2)，直到网络误差或迭代次数满足要求为止。

4　新型降雨预报模型的仿真和研究

本文研究的降雨预报模型是根据当天的气象数据预报第二天的降雨情况。降雨情况根据雨量标准可以被划分为以下级别：无雨(日降雨量为 0~0.1 mm)；小雨(0.1~9.9 mm)；中雨(10.0~24.9 mm)；大雨(25.0~49.9 mm)；暴雨(50.0~99.9 mm)；大暴雨(100.0~250.0 mm)；特大暴雨(超过 250.0 mm)。对于第二天降雨情况的预测级别和实际级别相一致，即可视为预报准确。降雨预报模型的预报准确率是评价它性能优劣的关键指标。

4.1　降雨预报模型的数据

本文选取山东省济南气象站(台站号：54823)2003~2006 年 8 月份的日值气象数据(124 条)作为研究对象，其中包括平均本站气压、最高本站气压、最低本站气压(0.1 hPa)、平均气温、最高气温、最低气温(0.1℃)、平均相对湿度、最小相对湿度(1%)、平均风速、最大风速、极大风速(0.1 m/s)、最大风速风向、极大风速风向(16 个方位)、日照时数(0.1 h)和第二日降水量(0.1 mm)15 项气象指标，如表 1 所示。

4.2　获得影响降雨的主成分和主成分数据集

本文采用 SPSS(Statistical Program for Social Sciences)统计软件以选取的日值气象数据为基础，对影响降雨的 14 个气象因素进行主成分分析。首先对该 14 个因素组成的数据集进行标准化，然后得到各因素间的相关矩阵。由相关矩阵求得其特征值并计算出各个主成分方差贡献率。本文选取方差累计贡献率超过 80%的前 4 个主成分，如表 2 所示。

表 3 给出了分别在这 4 个主成分上的负荷值，由式(4)计算出各主成分的系数向量，将系数向量与之前已标准化的数据集相乘，可以得到降维了的主成分数据集。表 4 给出了变换后的，与表 1 对应的主成分数据集(不含第二日降水量)。

4.3　样本预处理以及 BP 网络的设计

以上得到的主成分数据集，加上“第二日降水量”就是该降雨预报模型的样本数据集，其中主成分数据为输入数据，“第二日降水量”为理想输出数据。输入数据在使用前需要进行预处理，在这里按照

表1 日值气象数据(部分)

年	月	日	平均本站气压	最高本站气压	最低本站气压	平均气温	最高气温	最低气温	平均相对湿度	最小相对湿度	平均风速	最大风速	最大风速风向	极大风速	极大风速风向	日照时数	第二日降水量
2006	8	1	9 843	9 856	9 830	248	290	224	88	66	22	49	10	80	10	9	190
2006	8	2	9 873	9 885	9 846	262	320	227	86	55	14	39	5	52	4	45	336
2006	8	3	9 915	9 939	9 885	254	325	221	90	58	23	97	16	190	1	20	349
2006	8	4	9 923	9 947	9 910	249	310	214	88	59	17	42	5	62	13	12	3
2006	8	5	9 903	9 923	9 881	258	298	243	87	66	18	66	8	102	8	11	0
2006	8	6	9 878	9 897	9 856	262	320	240	82	55	28	69	8	94	6	73	100
2006	8	7	9 864	9 877	9 848	266	315	228	80	56	18	54	7	74	15	54	102
2006	8	8	9 871	9 880	9 857	273	321	239	81	58	25	54	14	78	13	69	0
2006	8	9	9 876	9 886	9 860	280	333	247	76	48	33	52	7	73	8	76	0
2006	8	10	9 876	9 890	9 860	286	338	256	69	45	26	49	7	67	7	87	0
2006	8	11	9 867	9 883	9 852	296	343	254	67	44	21	41	8	58	8	85	0
2006	8	12	9 852	9 866	9 834	301	359	248	72	43	26	39	7	52	8	77	0
2006	8	13	9 843	9 855	9 825	306	357	271	73	49	24	36	2	61	3	40	195
2006	8	14	9 839	9 848	9 821	282	320	257	89	70	24	55	3	86	3	13	4
2006	8	15	9 840	9 849	9 829	264	310	239	86	65	25	55	3	86	4	20	32 700
2006	8	16	9 853	9 862	9 836	273	317	232	69	46	29	58	5	93	5	106	0

注:数据来源于中国气象科学数据共享服务网;32 700 为特征值,代表微量的意思,在科学研究中可近似为 0。

表2 前4个主成分的方差和贡献率

主成分	方差	方差贡献率	方差累计贡献率
1	4.853	34.663	34.663
2	3.116	22.258	56.922
3	2.274	16.244	73.166
4	1.463	10.447	83.613

表3 各主成分的因子负荷量

	主成分			
	1	2	3	4
平均本站气压	−0.830	0.418	0.118	0.120
最高本站气压	−0.837	0.372	0.157	0.154
最低本站气压	−0.811	0.440	0.061	0.138
平均气温	0.860	0.266	−0.219	−0.131
最高气温	0.770	0.443	−0.242	−0.079
最低气温	0.837	−0.080	−0.084	−0.177
平均相对湿度	−0.122	−0.917	−0.013	0.104
最小相对湿度	−0.066	−0.903	0.141	0.079
平均风速	0.287	0.188	0.737	−0.071
最大风速	0.401	0.050	0.866	0.122
最大风速风向	0.357	0.044	−0.120	0.838
极大风速	0.402	0.039	0.844	0.159
极大风速风向	0.390	0.066	−0.273	0.765
日照时数	0.264	0.796	−0.065	−0.040

$$F_i[j]=\frac{F_i[j]-\min(F_i)}{\max(F_i)-\min(F_i)}\quad(i=1,2,3,4;j=1,\cdots,124)\qquad(10)$$

对输入数据进行归一化,使其值在 0~1 之间。

为了评价降雨预报模型的性能、准确率和泛化能力,本文从样本数据集中随机选取 70 条数据作为训练样本集,其余 54 条数据作为测试样本集。训练样本集中的“第二日降水量”将执行对应数据分别除以 10 000 的预处理办法。而测试样本集中的“第二天降水量”需要根据降雨级别划分标准进行分类。

该降雨预报模型所基于的 BP 网络的拓扑结构为:单隐层;输入层含 4 个节点;输出层含 1 个节点;根据“隐层节点数须小于训练样本数−1”和“训练样本数须多于连接权数,一般为 2~10 倍”的原则,确定隐层节点数为 6。连接权和节点阈值均初始化为 0~1 范围内的随机数,并取式(7)作为网络的误差度量函数,学习步长 η=0.6。

表4 与表1对应的主成分数据集

主成分			
1	2	3	4
0.951 8	−2.343 9	−0.942 4	0.663 5
0.025 6	−0.802 9	−1.859 9	−1.287 0
0.089 4	−0.215 7	2.890 5	1.574 6
−1.902 9	−0.285 1	−1.303 4	0.945 8
−0.429 5	−1.041 0	0.168 6	0.615 3
0.901 9	0.136 2	0.226 3	−0.211 3
1.141 8	−0.350 9	−1.435 2	0.844 1
1.580 5	−0.010 5	−1.167 3	1.735 1
1.356 5	0.897 4	−0.738 8	−0.521 3
1.448 2	1.498 5	−1.279 0	−0.825 8
1.775 1	1.500 1	−2.082 8	−0.674 6
2.392 3	1.050 3	−2.205 8	−1.055 4
2.461 1	0.024 4	−1.942 5	−2.860 3
1.869 1	−2.354 9	−0.509 8	−2.039 9
1.302 3	−1.985 6	−0.365 4	−1.739 3
1.703 1	1.020 8	−0.271 5	−1.418 7

4.4 BP 网络的改进效果对比

本文对于该降雨预报模型的仿真是在一台 PC 机上编程实现的。该 PC 机的具体性能如下:1.7 GHz 主频;256 MB 内存;40 GB 硬盘;安装并使用 Windows XP 操作系统;编程工具为 Microsoft Visual C++6.0。

在仿真过程中,首先利用训练样本集对降雨预报模型中的 BP 网络进行固定迭代次数(本文设定为 60 000 次)的训练,以学习降雨预报的内在规律。鉴于标准 BP 网络具有收敛速度慢和易陷于局部极小的弱点,仿真过程中采取以下措施,以提高网络的性能:采用加入动量项的权值调整规则进行网络训练(其中动量因子 α=0.9);由于 BP 网络具有“训练样本顺序敏感性”,本文采取在每次训练结束后检查各样本误差,并将误差最

大者再多训练一次的做法；隐层节点和输出层节点均采用 S 型函数 $f(x)=\frac{1}{1+e^{-2x}}$ 作为激活函数，其中 $f'(x)=2\cdot f(x)(1-f(x))$。

本文将以上三种改进措施组合起来用到 BP 网络的训练过程中，跟标准 BP 网络、加动量项的 BP 网络相比较，采用组合改进方案的 BP 网络取得了更好的训练效果。图 2 给出了三种 BP 网络在训练过程中的误差变化。

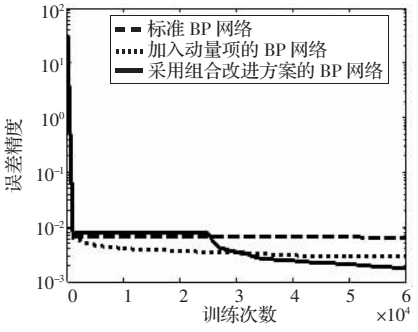

图 2　三种 BP 网络在训练过程中的误差变化

BP 网络训练过程结束后，需要利用测试样本集对该降雨预报模型的预报准确率进行评测。评测结果如图 3 所示，采用组合改进方案的 BP 网络的预报率明显高于标准 BP 网络和加动量项的 BP 网络。

5　结束语

影响降雨的天气因素有很多，而且各因素之间存在着内在联系和相关性。运用主成分分析法分析影响降雨的气象数据，不仅客观，更为重要的是在不牺牲数据准确性的情况下，减少了维数，从而在保证样本可靠性的前提下，为简化 BP 网络的拓扑结构，减少网络的训练时间奠定基础。本文研究的降雨预报模型所采用的改进 BP 网络，相比之下不仅训练效率更高，

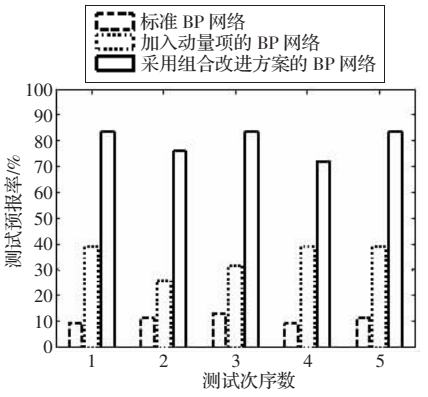

图 3　三种 BP 网络测试预报率对比

而且预报效果更好。

可见，主成分分析法和改进 BP 网络结合，是进行降雨预报的有效途径，特别适合解决像降雨预报这种没有现成理论公式依据，受多方面因素影响的多变量问题。不过，现实天气预报中类似“阵雨”、“雷阵雨”、“小雨转中雨”等天气状况的预测，利用主成分分析法和改进 BP 网络相结合还难以实现。此外，鉴于天气变化的滞后性，采用第一、二天的气象数据预测第三天的降雨量，也许会更客观、准确。

参考文献：

[1] Rumelhart D E，Hinton G E，Williams R J.Learning representations by back-propagation errors[J].Nature，1986，323：533-536.

[2] 郭陵之.应用于单站降雨预报的神经网络方法[J].计算机应用研究，2000，17(5)：28-30.

[3] 陶雪梅，刘自伟.基于神经网络的降雨预报系统及其改进[J].兵工自动化，2006，25(9)：60-65.

[4] 阎平凡，张长水.人工神经网络与模拟进化计算[M].北京：清华大学出版社，2000：17-26.

（上接 223 页）

mance enhancement and cross-talk reduction[C]//ICCAD-93，Digest of Technical Papers IEEE/ACM International Conference on Computer-Aided Design，7-11 Nov 1993：697-702.

[3] Chen H H，Wong C K.Wiring and crosstalk avoidance in multi-chip module design[C]//Proceedings of the IEEE Custom Integrated Circuits Conference，May 3-6，1992：28.6.1-28.6.4.

[4] Gao T，Liu C L.Minimum crosstalk switchbox routing[C]//IEEE/ACM International Conference on Computer-Aided Design，November 6-10，1994：610-615.

[5] Tong Gao，Liu C L.Minimum crosstalk channel routing[J].IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，1996，15(5)：465-474.

[6] Hameenanttila T，Carothers J D，Li D H.Fast coupled noise estimation for crosstalk avoidance in the MCG multichip module autorouter[J].IEEE Transactions on Very Large Scale Integration(VLSI) Systems，1996，4(3)：356-368.

[7] Kirkpatrick D A，Sangiovanni-Vincentelli A L.Techniques for crosstalk avoidance in the physical design of high-performance digital systems[C]//IEEE/ACM International Conference on Computer-Aided Design，November 6-10，1994：616-619.

[8] Stohr T，Alt M，Hetzel A，et al.Analysis，reduction and avoidance of crosstalk on VLSI chips[C]//1998 International Symposium on Physical Design，1988：211-218.

[9] Zhou H，Wong D F.Global routing with crosstalk constraints[J].IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，1999，18(11)：1683-1688.

[10] Tseng H P，Scheffer L，Sechen C.Timing-and crosstalk-driven area routing[J].IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，2001，20(4)：528-544.

[11] Kawaguchi H，Sakurai T.Delay and noise formulas for capacitively coupled distributed RC lines[C]//Proceedings of the ASP-DAC'98，Design Automation Conference，Asia and South Pacific，10-13 Feb，1998：35-43.

[12] Chang C C，Cong J.Pseudo pin assignment with crosstalk noise control[J].IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，2001，20(5)：598-611.

[13] Wu D，Hu J，Mahapatra R.Coupling aware timing optimization and antenna avoidance in layer assignment[C]//Proc ACM International Symposium on Physical Design，2005：20-27.

[14] Vittal A，Marek-Sadowska M.Reducing coupled noise during routing[C]//Proceedings，Fifth ACM SIGDA Physical Design Workshop，1996：27-33.