基于 PCA 的神经网络在港口货物吞吐量中的预测

穆俊鹏[1],李 娟[2],张 明[2]
(1.上海出版高等专科学校 信息化办公室,上海 200093;
2.上海海事大学 信息工程学院,上海 201306)

摘 要:分析选取了可能影响港口货物吞吐量的因素,采用 PCA 技术提取关键因子,最后以提取的关键因子作为神经网络的神经输入元,分别建立 BP 神经网络预测模型和 GRNN 神经网络预测模型。以上海港口为例,对港口货物吞吐量进行预测并对预测结果给予分析。

关键词:港口货物吞吐量;PCA;BP 神经网络;GRNN 神经网络;预测
中图分类号:TP391 文献标识码:A 文章编号:1674-7720(2012)19-0079-03

Harbor cargo throughput forecast based on BP neural network and PCA

Mu Junpeng[1], Li Juan[2], Zhang Ming[2]
(1.Shanghai Publishing and Printing College, Shanghai 200093, China;
2.Department of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

Abstract: This paper analyses the factors may influence Harbor cargo throughput, then use PCA technology extracted key factors. Finally, using key factors as the input artificial neural, respectively established the BP neural network predictive model and the GRNN neural network predictive model with which the model calculate and analyses results in the Shanghai cargo throughput forecast.

Key words: Harbor cargo throughput; PCA; BP neural network; GRNN neural network; forecast

随着经济的发展,港口的建设逐渐炽热化,但如果港口的发展规模供大于求就会造成资源的浪费[1],因此有必要对影响港口发展的港口货物吞吐量进行预测分析,使其成为港口建设和发展的理论依据。目前,在预测领域最为广泛的应用技术是人工神经网络模型,它具有较好的容错能力和较快的总体处理能力,而且能通过样本数据完成学习或训练,训练后的神经网络具有推广或者泛化能力(对未来数据的预测能力),故而本文采用神经网络模型对港口货物吞吐量进行预测。

1 研究方法

1.1 主成分分析法[2]

主成分分析也称主分量分析,它是将多指标转化为少数的几个综合指标,且这几个综合指标能够反映出原来多个变量反映的大部分信息。换言之,就是将许多相关性很高的变量转化成彼此相互独立的变量。

1.2 BP 神经网络

BP 神经网络的基本原理是:输入信号由输入层经过隐含层的变换函数作用后到达输出层得到目标信号,然后将输出的目标信号与实际数据相比较,利用输出后的误差来估计输出层的直接前导层的误差,再用这个误差估计更前一层的误差,如此一层一层地反传下去,就获得了所有其他各层的误差估计。但它存在一些缺点[3]:如:易陷入局部最小、收敛速度慢、隐含层的结点个数难以确定等问题。为了能够获得更好的泛化全局最优性能,主要完成以下的改进[4]:(1)提高网络训练的速度;(2)提高训练的精度;(3)避免网络陷入局部极小点。本文采用主成分分析法提取出关键因子主要是为了防止网络因影响因素过多而陷入局部极小点,从而提高模型的学习能力和泛化能力。

1.3 GRNN 神经网络[5]

广义回归神经网络是一种前馈式神经网络,不仅具有全局逼近的性能,还具有最佳逼近性能。它是依据非线性回归分析建立在非参数估计基础上的一种非线性回归径向基神经网络。由于 GRNN 的非线性映射能力较强,且网络最后收敛于样本聚集量较多的优化回归面,故常应用于函数逼近、模式分类等方面。

2 实例分析

选取 GDP、工业总产值、第一产业值、第二产业值、第三产业值、固定资产投资总额、进出口总额、社会消费品总额、人口总数、货运量、铁货物运输量和公货物运输量 12 个影响港口货物吞吐量的因素[2,6]。以上海港口为例收集了 12 个影响因素，数据来源于上海统计如表 1 所示。

2.1 利用 PCA 提取关键因子

步骤如下：

(1)将 12 个影响因素采用公式 $a[i,j]=v\dfrac{[i,j]}{\sum_{k=1}^{i}v[i,j]}$ 标准化矩阵，计算结果如表 2 所示。

(2)将上述矩阵依据公式 $r_{ij}=\dfrac{(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}$ 计算出影响因素之间的相关系数矩阵，其中 r_{ij} $(i,j=1,2,\cdots,p)$ 为原来变量 x_i 和 x_j 的相关系数。根据 R 得到 12 个影响因素之间的相关性。

(3)根据步骤(2)得出的矩阵求解对应的特征根和特征向量。首先求解特征方程 $|\lambda I-R|=0$ 的特征根 $\lambda_i(i=1,2,\cdots,p)$，并且按其大小顺序排列，即 $\lambda_1\geqslant\lambda_2\geqslant\cdots\lambda_p\geqslant0$；然后求出其对应的特征向量 $e_i(i=1,2,\cdots,p)$，结果如表 3 所示。

表 1 上海市吞吐量的 12 个影响因素

年份	吞吐量/万吨	GDP/亿元	工业总产值/亿元	第一产业值/亿元	第二产业值/亿元	第三产业值/亿元	固定资产投资总额/亿元	进出口总额/亿美元	社会消费品总额/亿元	人口数量/万人	货运量/万吨	铁货运量/万吨	公货运量/万吨
1995	16 567	2 499.43	4 547.47	59.82	1 419.41	1 020.20	1 601.79	190.25	1 050.96	1 301.37	22 531	1 376	6 273
1996	16 401	2 957.55	5 126.22	68.72	1 596.73	1 292.11	1 952.05	222.63	1 258.00	1 304.43	40 928	1 320	25 023
1997	16 397	3 438.79	5 649.93	72.03	1 744.02	1 592.74	1 977.59	247.64	1 435.38	1 305.46	41 373	1 252	25 991
1998	16 387	3 801.09	5 763.67	73.84	1 871.89	1 855.36	1 964.83	313.44	1 593.27	1 306.58	42 090	1 152	26 351
1999	18 641	4 188.73	6 213.24	74.49	1 984.64	2 12.960	1 856.72	386.04	1 722.33	1 313.12	44 485	997	27 171
2000	20 440	4 771.17	7 022.98	76.68	2 207.63	2 486.86	1 869.67	547.10	1 865.28	1 321.63	47 954	1 055	28 369
2001	22 099	5 210.12	7 806.18	78.00	2 403.18	2 728.94	1 994.73	608.98	2 016.37	1 327.14	49 545	1 080	28 869
2002	26 384	5 741.03	8 730.00	79.68	2 622.45	3 038.90	2 187.06	726.64	2 203.89	1 344.23	54 196	1 131	29 759
2003	31 621	6 694.23	11 708.49	81.02	3 209.02	3 404.19	2 452.11	1 123.97	2 404.45	1 341.77	58 669	1 208	30 678
2004	37 897	8 072.83	14 595.29	83.45	3 892.12	4 097.26	3 084.66	1 600.26	2 656.91	1 352.39	63 180	1 284	31 554
2005	44 317	9 247.66	16 876.78	90.26	4 381.20	4 776.20	3 542.55	1 863.65	2 979.50	1 360.26	68 741	1 278	32 684
2006	53 748	10 572.24	19 631.23	93.81	4 969.95	5 508.48	3 925.09	2 274.89	3 375.20	1 368.08	72 617	1 233	33 799
2007	56 144	12 494.01	23 108.63	101.84	5 571.06	6 821.11	4 458.61	2 829.73	3 873.30	1 378.86	78 108	1 143	35 634
2008	58 170	14 069.87	25 968.38	111.80	6 085.84	7 872.23	4 829.45	3 221.38	4 457.23	1 391.04	84 347	985	40 328
2009	59 205	15 046.45	24 888.08	113.82	6 001.78	8 930.85	5 273.33	2 777.31	5 173.24	1 400.70	76 967	941	37 745
2010	65 339	17 165.98	31 038.57	11.415	7 218.32	9 833.51	5 317.67	3 688.69	6 070.50	1 412.31	81 024	959	40 890

表 2 影响因素标准化结果

年份	GDP/亿元	工业总产值/亿元	第一产业值/亿元	第二产业值/亿元	第三产业值/亿元	固定资产投资总额/亿元	进出口总额/亿美元	社会消费品总额/亿元	人口数量/万人	货运量/万吨	铁货运量/万吨	公货运量/万吨
1995	0.019 8	0.020 8	0.043 6	0.024 8	0.015 1	0.033 2	0.008 4	0.023 7	0.060 5	0.024 3	0.074 8	0.013
1996	0.023 5	0.023 4	0.050 0	0.027 9	0.019 2	0.040 4	0.009 8	0.028 4	0.060 5	0.044 2	0.071 8	0.052
1997	0.027 3	0.025 8	0.052 4	0.031 0	0.023 6	0.041 0	0.010 9	0.032 4	0.060 7	0.044 6	0.068 1	0.054
1998	0.030 2	0.026 4	0.053 8	0.032 7	0.027 5	0.040 7	0.013 9	0.036 0	0.060 7	0.045 4	0.062 7	0.054 8
1999	0.033 3	0.028 4	0.054 2	0.034 7	0.031 6	0.038 5	0.017 1	0.038 9	0.061 0	0.048 0	0.055 4	0.056 5
2000	0.037 9	0.032 1	0.055 8	0.038 6	0.036 9	0.038 7	0.024 2	0.042 1	0.061 4	0.051 7	0.057 4	0.059
2001	0.041 4	0.035 7	0.056 8	0.042 0	0.040 5	0.041 3	0.026 9	0.045 6	0.061 7	0.053 5	0.058 7	0.06
2002	0.045 6	0.039 9	0.058 0	0.045 8	0.045 1	0.045 3	0.032 1	0.049 8	0.062 0	0.058 5	0.061 5	0.061 9
2003	0.051 3	0.053 5	0.059 0	0.056 1	0.050 5	0.050 8	0.049 7	0.054 3	0.062 4	0.063 3	0.065 7	0.063 8
2004	0.064 1	0.066 7	0.060 8	0.068 0	0.060 8	0.063 9	0.070 7	0.060 0	0.062 8	0.068 2	0.069 8	0.065 6
2005	0.073 4	0.077 2	0.065 7	0.076 6	0.070 9	0.073 4	0.082 4	0.067 3	0.063 2	0.074 2	0.069 5	0.067 9
2006	0.083 9	0.089 8	0.068 3	0.086 9	0.081 7	0.081 3	0.100 6	0.076 3	0.063 6	0.078 4	0.066 5	0.070 3
2007	0.099 2	0.105 7	0.074 2	0.097 4	0.101 2	0.092 3	0.125 1	0.087 5	0.064 1	0.084 3	0.062 2	0.074 1
2008	0.111 7	0.118 8	0.081 4	0.106 4	0.116 8	0.100 0	0.142 4	0.103 4	0.064 6	0.091 0	0.053 6	0.083 8
2009	0.119 4	0.113 8	0.082 9	0.104 9	0.132 5	0.109 2	0.122 8	0.116 9	0.065 1	0.083 0	0.051 2	0.078 5
2010	0.136 3	0.141 9	0.083 1	0.126 2	0.145 9	0.110 1	0.163 1	0.137 2	0.065 6	0.087 4	0.052 2	0.085

表 3 各成分的特征根,贡献率和累计贡献率

主成分	特征值	贡献率/%	累计贡献率/%
1	10.840 3	0.903 4	0.903 4
2	0.761 531	0.063 5	0.966 9
3	0.316 546	0.026 4	0.999 33
4	0.033 205	0.002 8	0.996 1
5	0.026 782 9	0.002 2	0.998 2
6	0.018 128 6	0.001 5	0.999 7
7	0.002 213 5	0.000 2	0.999 9
8	0.000 695	0.000 1	1
9	0.000 348 2	0	1
10	0.000 219 1	0	1
11	7.66E-5	0	1
12	-2.65E-16	0	1

(4) 根据公式 $\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_k}$ 和 $\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_k}$ 分别计算主成分的贡献率及累计贡献率,结果如表 3 所示。其中 $i=1,2,\cdots,p$,$\lambda_1,\lambda_2,\cdots,\lambda_m$ 所对应的是第 1、2,…,$m(m\leqslant p)$ 个主成分。

(5)由表 3 中可以看到成分 1、2 的累计贡献率达到 96%占主导地位,故舍弃其他的主成分保留主成分 1 和主成分 2,并根据公式 $l_{ij}=p(z_i,x_j)=\sqrt{\lambda_i}\,e_{ij}(i,j=1,2,\cdots,p)$ 计算主成分负荷,即 12 个影响因素分别在主成分 1、主成分 2 中所占的比重(单位为%),结果如表 4 所示。

表 4 主成分负荷

影响因素	主成分 1	主成分 2
GDP	0.995 1	-0.061 3
工业总产值	0.987 4	-0.136 3
第一产业值	0.994 5	0.055 5
第二产业值	0.990 1	-0.128
第三产业值	0.993 1	-0.016 8
固定资产投资总额	0.979 9	-0.139 3
进出口总额	0.982 6	-0.152 2
社会消费品总额	0.987 1	0.011 3
人口总额	0.990 2	-0.085 6
货运量	0.961 3	-0.040 2
铁货运量	-0.779 7	-0.779 7
公货运量	0.867	0.244 8

由表 4 中可以看出 GDP、第一、二、三产业值和人口总额在第一主成分中的负荷较大;铁货物运输量在第二主成分的负荷较大。因此得出关键因子为 GDP、第一、二、三产业值、人口总额和铁货物运输量。

2.2 神经网络

2.2.1 BP 网络结构

采用 Matlab 建立预测模型,输入神经元数为 6 个,隐含层为一层,激励函数为 tansig,隐含层的神经元个数根据比较法最终选取为 11 个(隐含层神经元个数为 11 个时,网络的预测值达到最佳),输出层的激励函数为 purelin,输出层的神经元个数为 1 个,即为预测的年港口货物吞吐量。

2.2.2 GRNN 网络结构

GRNN 网络结构是通过激活神经元来逼近函数,实现输入矢量的函数值由某一领域内的神经元矢量对应的函数值映射而逼近。结构如图 1 所示。

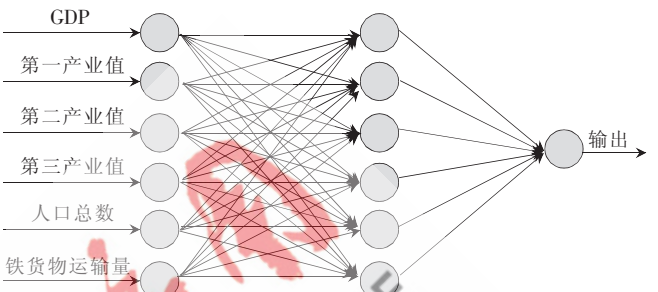

图 1 GRNN 网络结构模型

2.2.3 训练网络及其仿真结果

将前 15 组数据作为样本进行输入,对 BP 而言:采用 Levebberg-Marquardt 优化方法的训练函数 trainlm 对网络进行训练;权值的学习函数设为 learngdm,且学习率为 0.01;性能目标函数设为 mse;训练的次数为 1 000。对 GRNN 而言:输入层的神经元数等于输入样本数,其权值等于输入向量的转置 $IW=p^{\mathrm{T}}$,阈值 $b=[-\log(0.5)]^{1/2}/spread$,其中 spread 为径向基函数的扩展系数,此处扩展系数取值为 0.1(小一点的扩展系数可以更好地拟合数据);第二层神经元数也等于输入样本数,其目标向量为 T,无阈值向量,同样不需要训练;隐含层采用高斯变换来控制隐含层的输出,从而抑制输出单元的激活。

采用样本训练好的网络,以 2009 年和 2010 年的数据作为仿真数据进行测试,其测试结果如表 5 所示。

表 5 实验结果

年份	神经网络	实际值/万吨	预测值/万吨	误差/%
2009	BP	59 205	61 496	3.87
	GRNN	59 205	57 772	2.42
2010	BP	65 339	674 737	2.96
	GRNN	65 339	65 321	0.7

2.3 实验结果分析

从表 5 中可以看出,基于 PCA 的 BP 网络和 GRNN 网络在货物吞吐量预测中应用具有很好的拟合效果,相对误差都在 5%内,且基于 GRNN 网络的预测结果明显优于基于 BP 网络的预测结果。此外,还可以看到:(1)预测结果受到样本的大小的影响,当样本数量增多时,其预测效果更优;(2)PB 神经网络的预测模型具有不稳定性;(3)GRNN 神经网络的预测模型具有稳定性。

通过仿真得到的误差值验证了模型的可用性和模型的拟合能力。使用主成分分析有效地建立了互不相关的因子,减少了网络的复杂性,同时减小了落入局部区

域的概率;对于函数的拟合,BP网络表现出较好的拟合效果,而GRNN网络在预测方面的应用要优于BP网络。

但仍存在一定的缺陷:(1)未考虑突发因素的影响,只用于相对稳定的预测中;(2)对于BP网络中的隐含层,没有一种优异方法直接确定;(3)得到误差与预想的误差存在一定的界限。因此,本文下一步的目标是将神经网络与其他预测分析技术相结合,以便更好地减小误差。

参考文献

[1] 陈涛焘,高琴.港口集装箱吞吐量影响因素研究[J].武汉理工大学学报:信息与管理工程版,2008(6).

[2] 徐金河.基于主成分分析法的港口吞吐量内在影响因素研究[J].水运工程,2010(1).

[3] 王晨光,相秉仁,谢少斐,等.基于主成分分析的BP神经网络在药品销售预测中的应用[J].药物生物技术,2009(4).

[4] 龙训建,钱鞠,梁川.基于主成分分析的BP神经网络及其在需水预测中的应用[J].成都理工大学学报,2010(2).

[5] 王文才,王瑞智,孙宝雷,等.基于广义回归神经网络GRNN的矿井瓦斯含量预测[J].中国煤气层,2010,(1).

[6] 陈婷婷,陈漪翊.基于BP神经网络的港口货物吞吐量预测[J].计算机与现代化,2009(10).

(收稿日期:2012-04-16)

作者简介:

穆俊鹏,男,1970年生,工程师,本科,主要研究方向:计算机网络、校园信息化。

李娟,女,1986年生,硕士,主要研究方向:多媒体信息分析、处理与检索。

张明,男,1957年生,博士,教授,主要研究方向:多媒体应用技术、多媒体数据库、多媒体信息检索与处理、信息安全、物流、航运信息化技术。