·监测·

BP人工神经网络模型在上海市感染性腹泻日发病例数预测中的应用

黎健　顾君忠　毛盛华　肖文佳　金汇明　郑雅旭　王永明　胡家瑜

【摘要】 **目的**　建立基于气象因素的上海市感染性腹泻逐日发病例数BP人工神经网络预测模型。**方法**　收集上海市2005－2008年感染性腹泻逐日发病例数与同期气象资料包括气温、相对湿度、降雨量、气压、日照时数、风速，通过Spearman相关分析选出与感染性腹泻相关的气象因素，用主成分分析（PCA）去除气象因素间的共线性影响。利用MatLab R2012b软件的神经网络工具箱建立感染性腹泻日发病例数的BP神经网络预测模型，并对拟合效果、外推预测能力和等级预报效果进行评价。**结果**　Spearman相关性分析显示，日感染性腹泻与前两天的日最高气温、最低气温、平均气温、最低相对湿度、平均相对湿度呈正相关（$P<0.01$），与前两天的日平均气压呈负相关（$P<0.01$）。输入PCA提取的4个气象主成分构建BP神经网络预测模型，训练和预测样本平均绝对误差、均方根误差、相关系数、决定系数分别为4.7811、6.8921、0.7918、0.8418和5.8163、7.8062、0.7202、0.8180。模型预测值对2008年实际发病数的年平均误差率为5.30%，对感染性腹泻的等级预报正确率为95.63%。**结论**　温度和气压对感染性腹泻日发病例数影响较大。BP神经网络模型的拟合及预测误差较小，预报正确率较高，预报效果理想。

【关键词】 感染性腹泻；气象因素；BP人工神经网络

Preliminary application of Back-Propagation artificial neural network model on the prediction of infectious diarrhea incidence in Shanghai

LI Jian[1], *GU Jun-zhong*[2], *MAO Sheng-hua*[1], *XIAO Wen-jia*[1], *JIN Hui-ming*[1], *ZHENG Ya-xu*[1], *WANG Yong-ming*[2], *HU Jia-yu*[1]. 1 *Shanghai Municipal Center for Disease Control and Prevention, Shanghai* 200336, China; 2 *Institute of Computer Application, East China Normal University*

Corresponding authors: *WANG Yong-ming*, Email: ymwang@ica.stc.sh.cn; *HU Jia-yu*, Email: jyhu@scdc.sh.cn

This work was supported by a grant from the Constructing Program of Shanghai Municipal Public Health Key Discipline (No. 12GWZX0101).

【**Abstract**】 **Objective**　To establish BP artificial neural network predicting model regarding the daily cases of infectious diarrhea in Shanghai. **Methods**　Data regarding both the incidence of infectious diarrhea from 2005 to 2008 in Shanghai and meteorological factors including temperature, relative humidity, rainfall, atmospheric pressure, duration of sunshine and wind speed within the same periods were collected and analyzed with the MatLab R2012b software. Meteorological factors that were correlated with infectious diarrhea were screened by Spearman correlation analysis. Principal component analysis (PCA) was used to remove the multi-colinearities between meteorological factors. Back-Propagation (BP) neural network was employed to establish related prediction models regarding the daily infectious diarrhea incidence, using artificial neural networks toolbox. The established models were evaluated through the fitting, predicting and forecasting processes. **Results**　Data from Spearman correlation analysis indicated that the incidence of infectious diarrhea had a highly positive correlation with factors as daily maximum temperature, minimum temperature, average temperature, minimum relative humidity and average relative humidity in the previous two days ($P<0.01$), and a relatively high negative correlation with the daily average air pressure in the previous two days ($P<0.01$). Factors as mean absolute error, root mean square error, correlation coefficient (r), and the

DOI: 10.3760/cma.j.issn.0254-6450.2013.012.010

基金项目：上海市公共卫生重点学科计划（12GWZX0101）

作者单位：200336　上海市疾病预防控制中心（黎健、毛盛华、肖文佳、金汇明、郑雅旭、胡家瑜）；华东师范大学计算机应用研究所（顾君忠、王永明）

通信作者：王永明，Email：ymwang@ica.stc.sh.cn；胡家瑜，Email：jyhu@scdc.sh.cn

coefficient of determination (r^2) of BP neural network model were established under the input of 4 meteorological principal components, extracted by PCA and used for training and prediction. Then appeared to be 4.7811, 6.8921, 0.7918, 0.8418 and 5.8163, 7.8062, 0.7202, 0.8180, respectively. The rate on mean error regarding the predictive value to actual incidence in 2008 was 5.30% and the forecasting precision reached 95.63%. **Conclusion** Temperature and air pressure showed important impact on the incidence of infectious diarrhea. The BP neural network model had the advantages of low simulation forecasting errors and high forecasting hit rate that could ideally predict and forecast the effects on the incidence of infectious diarrhea.

【Key words】 Infectious diarrhea; Meteorological factors; Back-Propagation artificial neural network

全球每年约有30亿～50亿人发生感染性腹泻，死亡人数约为300万[1,2]。已有研究表明，感染性腹泻的发生、流行与气象因素密切相关[3-6]。感染性腹泻发病预测模型多基于统计学方法，纳入模型的气象因素相对较少，难以对气象因素与感染性腹泻发病之间非线性、多样化的关系进行有效描述[7-9]。BP人工神经网络已广泛应用于传染性疾病预测[10-12]。本研究收集上海市2005年1月至2008年12月感染性腹泻日发病数和同期气象资料，建立BP人工神经网络预测模型，并探讨其应用于医疗气象预报服务的可行性。

资料与方法

1. 资料：感染性腹泻日发病数据来源于国家疾病监测信息报告管理系统中2005年1月1日至2008年12月31日临床诊断或实验室确诊病例。同期上海地区主要气象资料由上海市气象局城市环境气象中心提供，包括日最高气温(℃)、最低气温(℃)、平均气温(℃)、最低相对湿度(%)、平均相对湿度(%)、平均气压(hPa)、降雨量(mm)、平均日照时数(hr)、平均风速(m/s)。

2. 气象主成分提取：考虑到气象因素之间存在共线性，根据主成分分析(PCA)原理[13,14]，应用MatLab R2012b软件对相关性分析得到的影响感染性腹泻发病的气象因素进行主成分提取，去除多重共线性。

3. 感染性腹泻日发病例数BP神经网络预测模型的建立：

(1)样本数据处理：2005－2007年的日气象数据和感染性腹泻日发病数为网络训练样本集(共1092对数据)，用于网络训练和权值修改。2008年的独立样本数据作为网络测试数据集(共366对数据)，用于检验模型的外推预测能力。为提高神经网络的训练速度和拟合效果，保证建立的模型具有良好外推能力，采用mapminmax函数对训练样本和测试样本进行归一化处理，并对预测结果进行反归一化处理。

(2)网络结构、参数设置及训练函数选择：采用MatLab R2012b软件中三层BP网络结构，以PCA提取的4个主成分作为网络输入(预测因子)，即输入层神经元数为4；以同期感染性腹泻日发病例数作为输出(预测项)，即输出层神经元数为1。在确定隐含层神经元个数时，通过经验公式[15]以及试错法发现当隐层神经元数为5时训练误差和测试误差最小。最后确定的神经网络结构为4-5-1，即4个输入节点，5个隐层节点，1个输出节点。为防止过拟合现象，不断调整网络参数，训练发现将网络最大训练次数设为1000次、训练步长为50、学习速率为0.01、动量系数为0.9、训练目标误差为0.001，可得最佳训练效果。用于网络建立、训练和仿真函数分别为newff、train和sim。

(3)模型拟合及预测效果检验：为评价BP神经网络模型的拟合和外推预测效果，采用平均绝对误差(MAE)、均方根误差(RMSE)、相关系数(r)及决定系数(r^2)等指标对所建的BP神经网络模型从训练拟合和外推预测两个方面进行检验。其中，MAE、RMSE值越小，r和r^2的值越接近1，说明预测准确度越高。首先，以2005—2007年逐日气象数据和同期感染性腹泻日发病例数对BP神经网络预测模型进行拟合效果检验；然后调用训练好的BP人工神经网络模型，以2008年的气象资料预测感染性腹泻发病数，计算模型的月相对误差百分率和年平均误差率。以预测结果的年平均误差率绝对值<20%为预测成功。

(4)模型等级预报效果检验：采用百分位数法，以2005—2008年感染性腹泻逐日发病例数的P_{50}、P_{75}、P_{95}三个值为预报阈值，将感染性腹泻日发病例数的预测值转换成对应的预报等级，进行腹泻指数等级预报。首先将预测得到的发病例数四舍五入取整，观察结果落在哪一级(表1)。如果预报等级与实际等级一致，则为预报命中；如果预报结果与实际等级一致或相差±1级，则为预报正确[8]，否则为预报失误。

表1　腹泻发病数预报等级划分

腹泻指数等级	模型预测值(P)	风险等级	公共卫生说明
一	$P<P_{50}$	低	不易发生腹泻(低发),注意饮食饮水卫生
二	$P_{50}\leqslant P<P_{75}$	中等	较易发生腹泻(易发),重视夏季饮食饮水卫生
三	$P_{75}\leqslant P<P_{95}$	高	易发生腹泻(易发),重视夏季饮食饮水卫生
四	$P\geqslant P_{95}$	极高	极易发生腹泻(极易发),高度重视夏季饮食饮水卫生

结　果

1. 感染性腹泻发病趋势:以散发为主,具有季节性发病高峰。每年的感染性腹泻逐日发病趋势基本一致,1—4月维持在较低水平,5月开始增加,7—8月达到全年峰值,然后回落,10—12月出现秋冬小高峰(图1)。

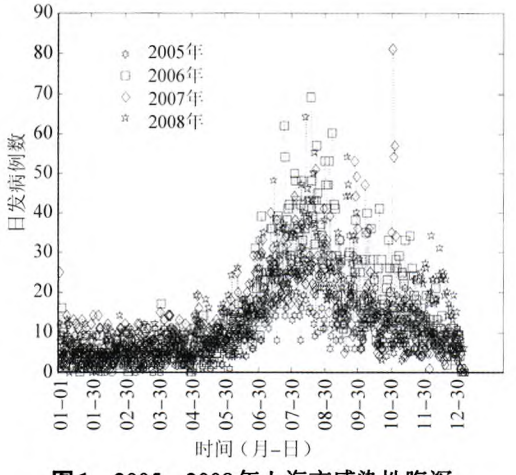

图1　2005—2008年上海市感染性腹泻日发病例数变化曲线

2. Spearman相关性分析:感染性腹泻的潜伏期一般为24～48 h。考虑到气象因素影响感染性腹泻发病可能存在滞后效应,选择感染性腹泻发病例数与当天、前一天、前两天的气象因素做Spearman相关性分析,结果显示,感染性腹泻日发病例数与当天、前一天、前两天的日最高、日最低及日平均气温均有较强的正相关性,且与前两天的日最高、最低及平均气温相关性最大,r分别为0.6284、0.6633和0.6590。与当天、前一天、前两天的日最低、平均相对湿度均呈正相关性,其中与前两天的日最低、平均相对湿度相关性最高,r分别为0.1063和0.0798。与当天、前一天、前两天的日平均气压均呈较高的负相关性,其中与前两天的日平均气压相关性最大($r=-0.4942$)。与当天、前一天、前两天的日照时数呈正相关性,其中与前一天的日照时数相关性最大($r=0.1300$)。与当天、前一天的日平均风速呈负相关,其中与当天的日平均风速相关性最大($r=-0.0646$)。上述r差异均有显著统计学意义($P<0.05$)。感染性腹泻发病与当天、前一天、前两天的日降雨量无明显相关性,r差异无统计学意义($P>0.05$),见表2。因此,使用前两天的日最高气温、最低气温、平均气温、最低相对湿度、平均相对湿度、平均气压、前一天的日照时数和当天的日平均风速8个气象因素作为候选预测因子,建立BP神经网络预测模型。

表2　感染性腹泻与气象因素的Spearman相关分析

气象因素(每日)	当日		前一天		前两天	
	r	P值	r	P值	r	P值
最高气温(X_1)	0.6259	<0.01	0.6253	<0.01	0.6284	<0.01
最低气温(X_2)	0.6559	<0.01	0.6540	<0.01	0.6633	<0.01
平均气温(X_3)	0.6544	<0.01	0.6548	<0.01	0.6590	<0.01
最低相对湿度(X_4)	0.0994	<0.01	0.0935	<0.01	0.1063	<0.01
平均相对湿度(X_5)	0.0768	<0.01	0.0710	<0.01	0.0798	<0.01
平均气压(X_6)	−0.4863	<0.01	−0.4902	<0.01	−0.4942	<0.01
降雨量(X_7)	−0.0073	>0.05	0.0163	>0.05	0.0201	>0.05
日照时数(X_8)	0.1278	<0.01	0.1300	<0.01	0.1094	<0.01
平均风速(X_9)	−0.0646	<0.05	−0.0645	<0.05	−0.0345	>0.05

3. 影响感染性腹泻气象因素的主成分提取:对原始输入样本进行预处理,求出其r矩阵。结果显示,日平均气温、最高气温、最低气温之间存在很大正相关性;以上三因素和平均气压之间存在很大负相关性;日最低相对湿度和平均相对湿度之间呈现很大正相关性,见表3,提示利用PCA去除多重共线性的必要性。

表3　r矩阵

气象因素	X_1	X_2	X_3	X_4	X_5	X_6	X_8	X_9
X_1	1.0000	0.9426	0.9797	0.0526	0.0685	−0.8646	0.1666	0.1540
X_2	0.9426	1.0000	0.9836	0.2660	0.2360	−0.8613	0.0937	0.1589
X_3	0.9797	0.9836	1.0000	0.1598	0.1499	−0.8743	0.1296	0.1592
X_4	0.0526	0.2660	0.1598	1.0000	0.8771	−0.2305	−0.2838	0.1030
X_5	0.0685	0.2360	0.1499	0.8771	1.0000	−0.2594	−0.2907	0.0919
X_6	−0.8646	−0.8613	−0.8743	−0.2305	−0.2594	1.0000	−0.0751	−0.1981
X_8	0.1666	0.0937	0.1296	−0.2838	−0.2907	−0.0751	1.0000	0.0384
X_9	−0.1540	0.1589	0.1592	0.1030	0.0919	−0.1981	0.0384	1.0000

计算矩阵r的特征值、主成分的方差贡献率和累积贡献率,并提取主成分。结果显示,前4个主成分包含原来4个指标全部信息的95.62%,因此选作BP神经网络的输入维数,见表4。

选定4个主成分后,计算主成分的特征向量矩阵,特征向量的绝对值越大,与主成分的相关性就越大,见表5。

表中PC1主要反映气温和气压的综合指标;PC2主要涵盖相对湿度信息;PC3、4主要反映风速和日照时数对感染性腹泻日发病的影响。因此,4

表4 各主成分的特征值和贡献率

气象因素	初始特征值			提取求和的平方载荷		
	特征值	各因素方差贡献率(%)	累计方差贡献率(%)	特征值	各因素方差贡献率(%)	累计方差贡献率(%)
X_1	3.8777	48.47	48.47	3.8777	48.47	48.47
X_2	2.0047	25.06	73.53	2.0047	25.06	73.53
X_3	0.9730	12.16	85.69	0.9730	12.16	85.69
X_4	0.7938	9.92	95.62	0.7938	9.92	95.62

表5 特征向量矩阵

气象因素	主成分(PC)			
	1	2	3	4
X_1	0.4836	−0.1658	−0.0805	−0.0689
X_2	0.4954	−0.0358	−0.0850	−0.0221
X_3	0.4955	−0.1033	−0.0843	−0.0501
X_4	0.1364	0.6344	0.0190	0.2742
X_5	0.1338	0.6384	0.0211	0.2526
X_6	−0.4693	0.0099	0.0296	0.0557
X_8	0.0678	−0.3823	0.2948	0.8726
X_9	0.1180	0.0655	0.9437	−0.2979

个主成分基本反映8个气象因素的信息。

4. 模型拟合及预测效果检验：以2005—2007年逐日气象数据和同期感染性腹泻日发病例数对BP神经网络预测模型进行拟合效果检验；同时，利用未参与模型拟合的2008年的独立样本数据作为测试样本数据对模型进行检验。

结果显示，BP神经网络预测模型的预测误差MAE、RMSE均在可接受的范围之内。训练样本和测试样本的r^2分别为0.841 80和0.818 00，说明拟合程度较好（表6）。利用2008年的气象数据对2008年感染性腹泻逐日发病数进行外推预测后按月汇总，计算2008年1—12月感染性腹泻的月发病实际值、预测值和月相对误差率。结果显示，模型预测值对2008年实际发病数的年平均误差率为5.30%，预测结果的平均误差＜20%，成功率为94.70%，见表7，提示模型实际应用效果理想。

表6 BP神经网络预测训练和测试样本性能指标

性能指标	训练样本	测试样本
MAE	4.781 10	5.816 30
RMSE	6.892 10	7.806 20
r	0.791 79	0.720 24
r^2	0.841 80	0.818 00

图2和图3分别显示BP神经网络预测模型对于训练样本和测试样本的预测值与实际值之间的趋势拟合（预测）曲线，拟合（预测）值曲线与实际值曲线非常接近，特别是在对高发期发病例数的预测上，BP神经网络预测模型获得理想效果。

表7 2008年1—12月感染性腹泻月发病预测数和月相对误差率

月份	实际发病数	预测发病数	相对误差率(%)
1	143	187	30.77
2	142	176	23.94
3	181	199	9.94
4	181	230	27.07
5	301	347	15.28
6	404	510	26.24
7	831	796	4.21
8	1023	877	14.27
9	764	685	10.34
10	386	481	24.61
11	434	327	24.65
12	508	202	60.23

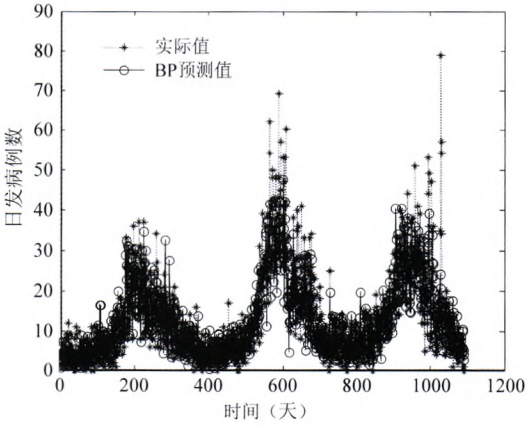

图2 BP网络模型训练样本拟合值（2005—2007年）与实际值拟合趋势曲线

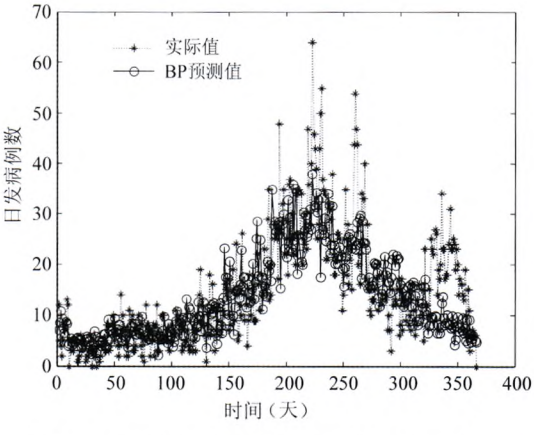

图3 BP网络模型预测样本预测值（2008年）与实际值拟合趋势曲线

5. 模型等级预报效果检验：为探讨模型应用于腹泻气象预报服务的可行性，利用2008年数据对模型做等级预报效果的检验。感染性腹泻日发病例数等级预报结果显示，BP神经网络预测模型有较高的预报正确率（95.63%），预报命中率（预报结果与实际等级完全一致）为58.47%，提示BP神经网络预测模型可应用于腹泻气象预报服务。

讨　论

感染性腹泻是受气象因素影响的传染性疾病[16]。气象因素与感染性腹泻发病之间一般为非线性关系，基于人工神经网络的预测模型可以不要求变量符合特定的分布类型，尤其善于处理模型复杂的非线性关系。模型不需要知道输入输出变量间的函数关系，只需通过对输入输出数据进行训练，获得输入、输出之间的映射关系，就能进行预测。

考虑到感染性腹泻具有一定的潜伏期，本研究对当天感染性腹泻发病例数与当天、前一天、前两天的气象因素进行相关分析，以消除气象因素对感染性腹泻发病的滞后效应影响。并且通过PCA去除气象因素之间的多重共线性影响，然后用PCA提取的4个气象因素主成分，建立上海市感染性腹泻日发病例数的BP神经网络预测模型，并使用参与模型拟合的训练样本数据和未参与模型建立的测试样本数据对所建的BP神经网络预测模型从拟合、预测和等级预报效果进行实验分析。结果表明，BP神经网络预测模型应用于感染性腹泻的预报具有较高的准确度，误差在合理范围之内，并且具有较好的等级预报能力，为感染性腹泻的预测预报提供了新方法，对于向公众发布腹泻气象指数预报有较好的应用价值。

影响感染性腹泻的气象因素较多，且各气象因素对感染性腹泻的影响程度不尽相同，因此，因素的选择应考虑到其对感染性腹泻的影响程度，充分考虑其相关性。气象因素与感染性腹泻之间以及气象因素之间存在着复杂的关系，如果不加取舍地全部纳入预测模型中，将影响模型预测效果。本研究显示，测试样本的预测效果略差于训练样本的拟合效果，说明拟合效果好并不代表有同样好的外推预测能力。因此，需要通过一定量的未参与模型建立的测试样本对所建立模型的预测能力进行检验，达到满意效果才能用于实际预测。极端天气变化也能影响感染性腹泻预测的准确度和预测模型的外推能力。此外，为使预测模型更加科学准确，有必要应用其他地区的数据进一步进行模型实际预测能力的检验。

参 考 文 献

[1] World Health Organization. Diarrhoeal disease. http://www.who.int/mediacentre/ factsheets/ fs330/en/index.html. [2013-4-15].

[2] Lin M, Dong BQ. Status in epidemiological research of infectious diarrhea. Chin Tropical Med, 2008, 8(4): 675-677. (in Chinese)
林玫，董柏青. 感染性腹泻流行病学研究现况. 中国热带医学，2008，8(4)：675-677.

[3] Lloyd SJ, Kovats RS, Armstrong BG. Global diarrhoea morbidity, weather and climate. Climate Res, 2007, 34(2): 119.

[4] Alexander KA, Carzolio M, Goodin D, et al. Climate change is likely to worsen the public health threat of diarrheal disease in Botswana. Internat J Environment Res Public Health, 2013, 10(4): 1202-1230.

[5] Singh RB, Hales S, de Wet N, et al. The influence of climate variation and change on diarrheal disease in the Pacific Islands. Environment Health Perspect, 2001, 109(2): 155.

[6] Kolstad EW, Johansson KA. Uncertainties associated with quantifying climate change impacts on human health: a case study for diarrhea. Environmental Health Perspect, 2011, 119(3): 299.

[7] Chou WC, Wu JL, Wang YC, et al. Modeling the impact of climate variability on diarrhea-associated diseases in Taiwan (1996-2007). Sci Total Environment, 2010, 409(1): 43-51.

[8] Zhao N, Ma XH, Gan L, et al. Research on the application of Medical-meteorological forecast model of infectious diarrhea disease in Beijing. IEEE Fifth International Conference, 2010.

[9] Hashizume M, Armstrong B, Hajat S, et al. Association between climate variability and hospital visits for non-cholera diarrhoea in Bangladesh: effects and vulnerable groups. Internat J Epidemiol, 2007, 36(5): 1030-1037.

[10] Yu B, Ding C, Wei SB, et al. Early warning on measles through the neural networks. Chin J Epidemiol, 2011, 32(1): 73-76. (in Chinese)
余滨，丁春，魏善波，等. 神经网络在麻疹预测预警中的应用. 中华流行病学杂志，2011，32(1)：73-76.

[11] Gao HL, Lan L, Qiao DJ, et al. A preliminary study on the effects of meteorological factors on intracerebral hemorrhage death using the BP neural network model. Chin J Epidemiol, 2012, 33(9): 937-940. (in Chinese)
高菡璐，兰莉，乔冬菊，等. BP神经网络模型用于气象因素对脑出血死亡影响的初步研究. 中华流行病学杂志，2012，33(9)：937-940.

[12] Xu JF, Zhou XN. Application of artificial neural networks in infectious diseases. Chin J Parasitol Parasitic Dis, 2011, 29(1): 49-54. (in Chinese)
徐俊芳，周晓农. 人工神经网络在传染病研究中的应用. 中国寄生虫学与寄生虫病杂志，2011，29(1)：49-54.

[13] Fan JC, Mei GL. Data Analysis. Beijing: Science Press, 2002: 141-154. (in Chinese)
范金城，梅良林. 数据分析. 北京：科学出版社，2002：141-154.

[14] Wang XR, Wang SG. Practical Multivariate Statistical Analysis. Shanghai: Shanghai Scientific and Technical Publishers, 1990: 270-344. (in Chinese)
王学仁，王松桂. 实用多元统计分析. 上海：上海科学技术出版社，1990：270-344.

[15] Pei Z. Thinking of BP artificial neural network hidden layer structure design. Chin Electronic Commerce, 2011(10): 44-45.
裴志. BP人工神经网络隐层结构设计的思考. 中国电子商务，2011(10)：44-45.

[16] Wang DS. 2011 the Yaohai district of the district of infections diarrhea incidence of temporal and spatial distribution characteristics and analysis of the telated factor. J Med Theor Prac, 2012, 25(11): 1300-1301, 1312. (in Chinese)
王大顺. 2011年瑶海区感染性腹泻发病时空分布特征及相关影响因素分析. 医学理论与实践，2012，25(11)：1300-1301，1312.

（收稿日期：2013-07-15）
（本文编辑：万玉立）