

线性回归与 BP 神经网络方法的山东粮食产量预测比较研究

高亮亮^{1,2},潘彩霞^{1,3},屠星月^{1,3} (1. 农业部农业信息获取技术重点实验室,北京 100083;2. 山东农业大学信息学院,山东泰安 271018;3. 中国农业大学信息与电气工程学院,北京 100083)

摘要 山东省是我国传统农业大省,粮食产量对我国粮食总产量的影响较大,因此对山东省粮食产量进行预测具有重大意义。分别利用多元线性回归方法和 BP 神经网络两种预测方法对山东粮食产量进行预测,并对两种方法的预测结果进行分析比较,实验证明,BP 神经网络平均预测精度高于多元线性回归模型,且各期预测精度较多元线性回归模型更稳定,但随时间推移,误差增大,因此 BP 神经网络预测模型较适用于近期粮食产量预测。
关键词 线性回归;BP 神经网络;粮食产量预测
中图分类号 S126 **文献标识码** A **文章编号** 0517-6611(2014)30-10780-04

Comparison and Research of Linear Regression and BPNN for Shandong Province Grain Production Prediction
GAO Liang-liang^{1,2}, PAN Cai-xia^{1,3}, TU Xing-yue^{1,3} (1. Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture, Beijing 100083; 2. College of Information Science and Engineering, Shandong Agricultural University, Tai'an, Shandong 271018; 3. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083)
Abstract Shandong Province is a traditional agricultural province, of which grain production accounts for a high share of China's grain production. Therefore, grain production prediction of Shandong Province is of great importance. In this paper, linear regression and BP neural network are adopted for grain production forecast of Shandong Province, and a result compare between the two models is analyzed showing that BP neural network achieves higher average accuracy than the linear regression, and keeps more stable period accuracy.
Key words Linear regression; BP neural network; Grain production prediction

粮食是人类赖以生存的必要条件,是从事其他社会事务的前条件,也是关系到社会稳定、国家经济发展的重大战略问题,被视为当前世界面临的重大问题之一^[1]。由于极端气候变化频发、耕地面积减少、从事农业的劳动人员减少等因素对粮食生产的影响,我国粮食产量波动频繁且幅度较大,因此科学准确的粮食产量预测对我国粮食安全有着重大意义。山东省是我国农业大省,2013 年粮食产量仅次于黑龙江省,位居全国粮食产量第 2 名^[2],因此,笔者以山东省粮食产量为例进行预测方法比较研究。

粮食产量预测一般选择与粮食产量相关联的影响因素,依据特定方法理论确定相关关系,从而进行产量预测。目前,常用的方法包括传统的统计学预测方法、灰色系统预测方法以及支持向量机、神经网络等人工智能预测方法等^[3-10]。为此,笔者选择了两类常用的预测方法,线性回归方法和 BP(Back Propagation)神经网络,对山东省粮食产量进行预测,并分析评价两种方法的预测结果,确定适用于山东省粮食产量预测的方法。

1 研究区域与数据选择

1.1 山东省地区概况 山东省位于我国东部沿海的东北段,面积 15.71 万 km²,东邻黄海,北接渤海,内陆与河北、河南、安徽和江苏接壤。海拔较低,地形多样,平原占 55.0%,山地占 15.0%,丘陵占 20.0%,湖泊占 0.8%,其余占 9.3%。气候为典型的温带大陆性季风气候,降水集中,雨热同期,春季较短,气候多变,降水量较少;夏季炎热,多降雨;秋天余量较少,秋高气爽;冬季容易受到北方寒流的影响,寒冷干燥。

山东是我国农业大省,农业历史悠久,传统粮食作物有小麦、玉米、红薯、大豆、水稻、粳稻等,2013 年粮食产量达到 4 528.2 万 t,位居全国粮食产量第 2。

1.2 数据来源 影响粮食产量的因素很多,主要分为自然因素、劳动力因素、农用水利设施因素、化肥施用因素和农机电因素等。该研究以中华人民共和国国家统计局 1978~2011 年山东省粮食产量及粮食作物播种面积、有效灌溉面积、农用化肥施用折纯量、农村用电量、受灾面积、成灾面积、农业机械总动力、农林牧渔业从业人员 8 个影响因素的年度数据为例进行研究分析,详情如表 1 所示,粮食产量变化趋势如图 1 所示。可见,随着时间变化,粮食产量呈上下波动,并存在整体上升的趋势。

1.3 影响因素选择 为分析各影响因素与粮食产量的相关性,确定相关性较大的影响因素,避免相关性较小因素对预测结果的干扰,对影响因素进行了关联度分析。由于各个变量之间量纲不统一,需要进行归一化处理,设上述自变量中粮食作物播种面积为 X_1 、有效灌溉面积为 X_2 、农用化肥施用折纯量为 X_3 、农村用电量为 X_4 、受灾面积为 X_5 、成灾面积为 X_6 、农业机械总动力为 X_7 、农林牧渔业从业人员为 X_8 ,粮食产量为因变量 Y 。归一化公式为:

$$\hat{x}_i = \frac{x_i - \min X_i}{\max X_i - \min X_i} (i = 1, 2, \dots, 8)$$

(1)

式中, $\hat{x}_i$ 为归一化之后的值; x_i 为自变量 X_i 当前值; $\max X_i$ 、 $\min X_i$ 为该列中最大数值和最小数值。因变量 Y 也用同样方法进行归一化处理。归一化处理之后,根据灰色关联度进行自变量的关联度分析,公式为:

$$\xi_i(k) = \frac{\min \min |x_0(k) - x_i(k)| + \rho \max \max |x_0(k) - x_i(k)|}{|x_0(k) - x_i(k)| + \rho \max \max |x_0(k) - x_i(k)|}$$

(2)

基金项目 山东省自主创新专项(2012CX90204)。
作者简介 高亮亮(1990-),男,山东潍坊人,硕士研究生,研究方向:农业信息化。
收稿日期 2014-09-16

表 1 1990 ~ 2011 年山东省粮食产量及其影响因素

年份	粮食作物播种 面积// km^2	有效灌溉面积 km^2	农用化肥施 用折纯量//万 t	农村用电量 亿 $\text{kW} \cdot \text{h}$	受灾面积 km^2	成灾面积 km^2	农业机械总 动力//万 kW	农林牧渔业从 业人员//万人	粮食产量 万 t
1990	8 151.90	4 463.67	245.50	75.70	2 849.30	1 430.00	3 215.80	2 507.90	3 354.90
1991	8 088.10	4 552.33	271.50	84.20	2 414.70	1 208.00	3 304.70	2 647.20	3 916.80
1992	7 918.60	4 596.70	281.90	100.20	5 008.00	2 665.00	3 374.70	2 627.40	3 589.30
1993	8 213.40	4 624.00	355.00	106.60	3 743.00	1 933.00	3 518.00	2 597.00	3 964.60
1994	8 014.10	4 642.00	326.60	132.33	3 602.00	1 345.00	3 756.40	2 529.60	3 921.90
1995	8 131.60	4 662.50	362.30	147.35	2 331.00	741.00	4 016.47	2 502.20	4 246.40
1996	8 237.29	4 692.70	369.40	152.43	4 168.00	1 464.00	4 308.90	2 475.10	4 332.70
1997	8 083.13	4 736.70	386.70	169.46	4 746.00	2 717.00	4 763.60	2 497.54	3 852.22
1998	8 132.52	4 780.19	406.54	168.23	1 541.00	510.00	5 228.32	2 487.00	4 264.83
1999	8 099.25	4 805.52	419.28	173.83	3 000.00	1 230.00	6 096.58	2 474.13	4 269.00
2000	7 363.19	4 824.86	423.19	200.27	3 697.00	2 271.00	7 025.24	2 462.62	3 837.74
2001	7 153.51	4 836.10	428.62	214.05	3 596.00	2 281.00	7 689.57	2 434.28	3 720.60
2002	6 912.61	4 797.44	433.92	238.28	4 941.90	3 179.40	8 155.63	2 370.91	3 292.69
2003	6 415.41	4 760.79	432.65	272.24	2 632.00	1 250.00	8 336.70	2 264.62	3 435.54
2004	6 176.32	4 766.81	450.96	304.14	2 119.10	771.00	8 751.90	2 180.12	3 516.70
2005	6 711.73	4 789.96	467.63	346.54	1 779.80	712.60	9 199.33	2 045.93	3 917.38
2006	6 999.13	4 818.16	489.82	376.17	1 880.10	1 105.30	9 555.28	2 011.82	4 092.97
2007	6 936.49	4 836.78	500.34	408.18	1 865.40	636.20	9 917.79	1 949.98	4 148.76
2008	6 955.61	4 857.48	476.33	400.01	672.30	233.80	10 350.00	1 991.87	4 260.50
2009	7 030.09	4 896.92	472.86	415.23	2 341.90	1 182.30	11 080.66	1 984.42	4 316.30
2010	7 084.80	4 955.30	475.32	439.03	2 582.30	1 052.00	11 628.97	1 993.42	4 335.68
2011	7 145.82	4 986.88	473.64	456.52	2 117.20	415.60	12 098.25	1 981.21	4 426.29

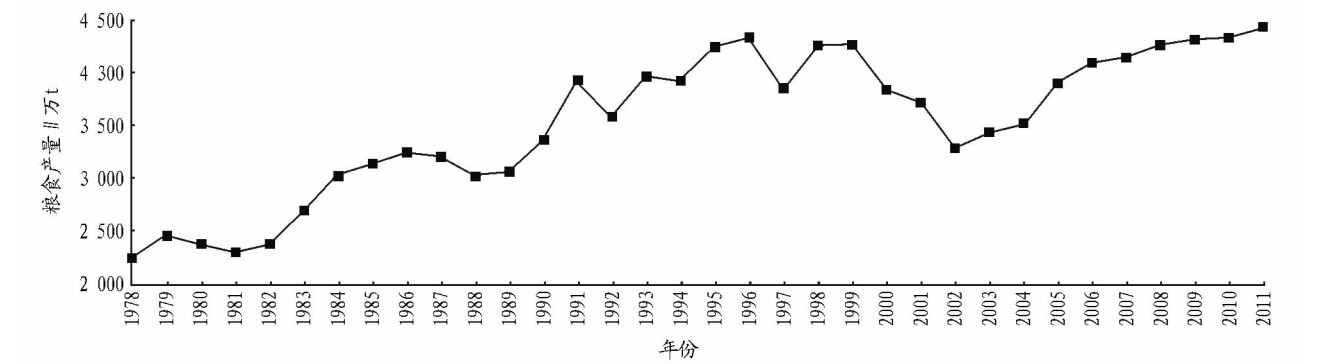


图 1 山东省粮食产量(1990 ~ 2013)走势

式中, ρ 称为分辨系数。 ρ 越小,分辨力越大。 $0 < \rho \leq 1, \rho$ 取 0.5。由于归一化之后双极小值为 0,双极大值为 1,所以该式可简化为:

$$\xi_i(k) = \frac{0.5}{|y_i(k) - x_i(k)| + 0.5}$$

(3)

比较序列对参考序列的灰色关联度常记为 $r = (y_i, x_i)$ 。常用的计算关联度的方法为平均值法:

$$r = (x_0, x_i) = \frac{1}{n} \sum_{k=1}^n \xi_i(k)$$

(4)

$r = (y_i, x_i)$ 为第 i 个影响因子序列关于粮食产量序列的关联度,依据 $r = (y_i, x_i)$ 值的大小可确定各个影响因子序列对于粮食产量序列的关联程度,对数据进行关联度分析,计算结果如表 2 所示。由表 2 可知,对粮食产量影响由高到低分别为农用化肥使用折纯量、有效灌溉面积、农业机械总动力、农业用电量、农林牧渔从业人员、粮食播种面积、成灾面积、受灾面积。这说明,农药化肥和灌溉面积对粮食产量的影响较大,由于山东的地理位置和气候条件,遭受自然灾害相对较轻,因此受灾面积、成灾面积对粮食产量有影响但相对较轻。

表 2 各个因素对粮食产量关联度分析结果

影响因素	影响因子(r)	排序
有效灌溉面积	0.77	2
农用化肥施用折纯量	0.81	1
农村用电量	0.73	4
受灾面积	0.60	8
成灾面积	0.60	7
农业机械总动力	0.73	3
粮食作物播种面积	0.60	6
农林牧渔业从业人员	0.62	5

2 粮食产量预测模型应用与对比

采用传统方法多元线性回归和人工智能方法 BP 神经网络对山东省粮食产量进行预测,并对其结果进行比较分析。

2.1 多元线性回归预测 多元线性回归方法是一种根据现有数据发掘变量之间相互关系的统计学方法,是根据现有的数据构建线性回归模型,达到对未来趋势作出预测的目的。多元回归方程如下:

$$Y = c_0 + c_1X_1 + c_2X_2 + c_3X_3 + c_4X_4 + c_5X_5 + c_6X_6 + c_7X_7 + c_8X_8$$

(5)

式中,因变量 Y 是粮食产量; c_0 是误差项; X_1 、 X_2 、 $\cdots$ 、 X_8 分别为粮食作物播种面积、有效灌溉面积、农用化肥施用折纯量、农村用电量、受灾面积、成灾面积、农业机械总动力、农林牧渔业从业人员; c_1 、 $\cdots$ 、 c_8 为对应回归系数。以 1978~2006 年 8 个因素 232 个样本为训练集,以 2007~2011 年数据作为检验集,利用软件 SPSS 12.0 进行数据统计分析,采用 Enter 进行分析,由表 3 可知, $Sig. = 0 < 0.005$,因此该模型具有显著统计学意义。

表 3 显著性检验结果

模型	平方和	自由度	均方	F 统计量	Sig.
回归	10 871 209.34	8	1 358 901.17	32.92	0
残差	825 671.86	20	41 283.59		
合计	11 696 881.20	28			

由表 4 可知,回归方程常数项为 -5 524.734,粮食作物播种面积回归系数为 0.483,有效灌溉面积回归系数为 0.471,农用化肥施用折纯量回归系数为 5.575,农村用电量回归系数为 0.195,受灾面积回归系数为 0.146,成灾面积回归系数为 -0.429,农业机械总动力回归系数为 -0.005,农林牧渔业从业人员回归系数为 0.638,因此回归方程为:

$$Y = 0.48X_1 + 0.471X_2 + 5.58X_3 + 0.20X_4 + 0.15X_5 - 0.43X_6 + 0.01X_7 + 0.64X_8 - 5\,524.73$$

(6)

表 4 系数求解结果

模型	非标准化系数		标准化系数	t 统计量	Sig.
	B	标准误差	Beta		
常量	-5 524.734	4 856.060		-1.140	0.270
农业机械总动力	-0.005	0.115	-0.019	-0.041	0.968
农村用电量	0.195	3.227	0.031	0.061	0.952
农林牧渔业从 业人员	0.638	0.917	0.146	0.696	0.495
农用化肥施用折 纯量	5.575	3.031	1.072	1.840	0.081
受灾面积	0.146	0.111	0.224	1.313	0.204
成灾面积	-0.429	0.177	-0.428	-2.424	0.025
有效灌溉面积	0.471	0.722	0.117	0.653	0.521
粮食作物播种 面积	0.483	0.145	0.490	3.332	0.003

2.2 BP 神经网络 BP 神经网络于 1986 年由 Rumelhart 和 McClland 为首的科学家小组提出,是一种按误差逆传播算法训练的多层前馈网络,是目前应用最广泛的神经网络模型之一。BP 神经网络分为 3 层:输入层、隐藏层和输出层。输入层为数据作为变量输入网络中,输出层为输出结果,包含 1 个或多个神经元,两者之间为隐藏层,隐藏层可以为 1 个或多个。

BP 神经网络的学习训练过程由网络输入信号正向传播和误差信号反向传播两部分组成。在正向传播中,输入信息从输入层经隐含层逐层计算传向输出层,输出层的各神经元输出对应输入模式的网络响应;如果输出层得不到期望输出,则误差转入反向传播,按减小期望输出与实际输出的误差原则,从输出层经中间各层,最后回到输入,层层修正各个

连接权值。随着这种误差逆传播训练不断进行,网络对输入模式响应的正确率也不断提高,如此循环直到误差信号达到允许的范围之内或训练次数达到预先设计的次数为止。

BP 神经网络学习规则是对网络权值和阈值的修正。
要沿着表现函数下降最快的方向——负梯度方向:

$$x_k + 1 = X_k - a_k g_k$$

(7)

式中, x_k 为当前的权值和阈值矩阵; g_k 为当前表现函数的梯度; a_k 为学习速率。

对 3 层 BP 网络,输入节点为 x_i ,隐层节点为 y_j ,输出节点为 z_l ,输入节点与隐层节点的网络权值为 w_{ji} ,隐层节点与输出节点的网络权值为 w_{li} ,当输出节点的期望为 t_l 时,模型的计算公式如下。

隐层节点的输出:

$$y_j = f(\sum_i w_{ji} x_i - \theta_j)$$

(8)

输出节点的计算输出:

$$z_l = f(\sum_j v_{lj} y_j - \theta_l)$$

(9)

输出节点误差:

$$E = \frac{1}{2} \sum_l (t_l - z_l)^2 = \frac{1}{2} \sum_l \{ t_l - f[\sum_j v_{lj} f(\sum_i w_{ji} x_i - \theta_j) - \theta_l] \}^2$$

(10)

如图 2 所示,该研究中输入层为影响粮食产量的 8 个因素,输出层为粮食产量。以 1978~2006 年 8 个因素 232 个样本为训练集,以 2007~2011 年数据作为检验集,利用软件 SPSS 12.0 进行 BP 神经网络建模。其网络参数设置如下:①网络节点,输入层因子为 8,输出层节点数为 1。②初始权值的确定,在程序中设计了一个随机发生器程序,产生 1 组 -0.5~0.5 的随机数,作为网络的初始权值。③使用方法为 Quick,设定模拟时间 5 min,使用 expert 建模,选择隐藏层为 2 层,第 1 层存在 20 个神经元,第 2 层存在 10 个神经元。④迭代次数为 500 次。

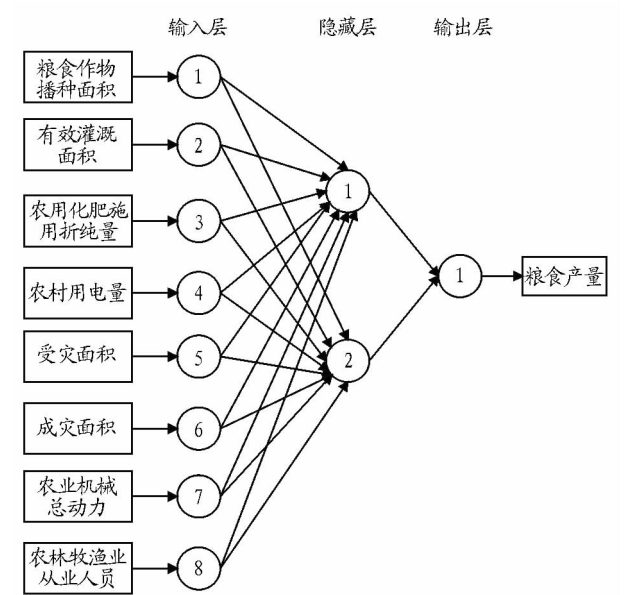


图 2 神经网络示意

2.3 预测结果比较分析 通过上述多元线性回归方法和

BP 神经网络建立了粮食产量预测模型,以 1978 ~ 2006 年 8 个因素 232 个样本数据作为训练集构建模型,以 2007 ~ 2011 年数据作为检测集对模型进行检测,检测结果如表 4 所示。由表 4 可知,BP 神经网络整体预测精度高于多元线性回归模型。基于多元线性回归方法的各年粮食产量预测误差差

异较大,相对误差 0.50% ~ 8.66%,预测效果不平稳;基于 BP 神经网络的粮食产量预测精度变化较小,除 2007 年外,其他年度精度均在 2.62% ~ 4.06%,预测效果较稳定,但是预测模型随着预测年数的增长呈现误差增大的现象,因此该模型适用于未来两期粮食产量预测。

表 5 粮食产量预测结果对比

年份	实际值 万 t	多元线性回归方法			BP 神经网络方法		
		预测值//万 t	绝对误差//万 t	相对误差//%	预测值//万 t	绝对误差//万 t	相对误差//%
2007	4 148.76	4 169.56	20.80	0.50	4 162.37	13.61	0.33
2008	4 260.50	4 076.68	183.82	4.31	4 148.77	111.73	2.62
2009	4 316.30	3 942.71	373.59	8.66	4 148.76	167.54	3.88
2010	4 335.68	4 109.18	226.50	5.22	4 164.28	171.40	3.95
2011	4 426.29	4 343.12	83.17	1.88	4 246.46	179.8	4.06
平均绝对误差			177.58			128.82	
平均相对误差				4.11			2.97

3 结论

山东省作为我国粮食产量大省,粮食产量对我国粮食安全有着重要的影响,所以对山东省粮食产量预测有着重要意义。为此该研究利用多元线性回归和 BP 神经网络两种方法对山东省粮食产量进行了预测,并对预测结果进行对比分析,结果表明:BP 神经网络平均预测精度高于多元线性回归模型,且各期预测精度较多元线性回归模型更稳定,但随时间推移,误差增大,因此 BP 神经网络预测模型较适用于近期粮食产量预测,例如近两年产量预测,若需对多期未来粮食产量进行预测,则需对 BP 神经网络算法改进,或对参数设置进行改进进一步研究。

参考文献

[1] 李晓东. 基于灰色关联支持向量机的中国粮食产量预测模型[J]. 河北理工大学学报:自然科学版,2008(4):76-80,103.

[2] 国家统计局. 国家统计局关于 2013 年粮食产量的公告[EB/OL]. (2013-11-29) [http://www. stats. gov. cn/tjsj/zxfb/201311/t20131129_475486. html](http://www.stats.gov.cn/tjsj/zxfb/201311/t20131129_475486.html).
[3] 周永生,肖玉欢,黄润生. 基于多元线性回归的广西粮食产量预测[J]. 南方农业学报,2011(9):1165-1167.
[4] 林芳. 灰色神经网络在粮食产量预测中的应用[J]. 计算机仿真,2012(4):225-228,267.
[5] 宰松梅,温季,仵峰,等. 基于灰色关联分析的辽宁省粮食产量预测模型[J]. 节水灌溉,2011(5):64-66.
[6] 旷岭. RBF 神经网络的粮食产量预测[J]. 计算机仿真,2011(11):189-191,200.
[7] 向昌盛,张林峰. 灰色理论和马尔可夫相融合的粮食产量预测模型[J]. 计算机科学,2013(2):245-248.
[8] 孙东升,梁仕莹. 我国粮食产量预测的时间序列模型与应用研究[J]. 农业技术经济,2010(3):97-106.
[9] 向昌盛,周子英,武丽娜. 粮食产量预测的支持向量机模型研究[J]. 湖南农业大学学报:社会科学版,2010(1):6-10.
[10] 李涛. 基于遗传神经网络的粮食产量预测方法研究[D]. 哈尔滨:哈尔滨工程大学,2008.

(上接第 10761 页)

的第一步,使农民工转化为“市民”是提高其收入与保障的关键点。第三,完善农副产品的补贴机制。随着我国经济水平的快速提高,市场对农副产品的需求量越来越大,扩大了农副产品生产的规模,很大一部分农民加入到农副产品生产行业,如各类养殖业、保健植物种植业等。农副产业的崛起快速,其扶助、保障政策还不够完善,农民收入受不确定因素影响严重,因此,制定和完善农副产业的各类政策,是保障农民收入稳步提升的另外一个关键。

参考文献

[1] 郑素芳,张岳恒. 我国农民收入的影响因素及对策建议[J]. 广东农业科学,2010(10):223-225.
[2] 宋委倩. 诸城市农民收入增长问题研究[D]. 青岛:中国海洋大学,2010:40-49.
[3] 卢立香. 中国金融发展对农民收入增长影响的理论与实证研究[D]. 济南:山东大学,2009:102-105.
[4] 王亚娜,查奇芬,宋丽萍. 影响农民收入因素的实证分析[J]. 安徽农业科学,2007(1):262-263.
[5] 郭庆春,何振芳,李力,等. 基于 BP 神经网络的我国农民收入预测模型[J]. 安徽农业科学,2011,39(20):12471-12473.
[6] BOX Q E P,JENKINS G M. Times series analysis:forecasting and control

[M]. San Francisco:Holden-day Press,1970:40-185.
[7] HANNAN E J. The estimation of the order of an ARMA process[J]. Ann Statist,1980,8(5):1071-1081.
[8] 邓自立,郭一新. 动态系统分析及其应用[M]. 沈阳:辽宁科学技术出版社,1985:31-130.
[9] MAY R,DANDY G C,MAIER H R,et al. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems[J]. Environmental Modeling & Software,2008,3(8):1-11.
[10] 何燕,何慧,孟翠丽,等. 基于 BP 人工神经网络方法的广西飞虱发生等级预测[J]. 生态学杂志,2013,33(1):1-10.
[11] 袁哲明,张永生,熊洁仪,等. 基于 SVR 的多维时间序列分析及其在农业科学中的应用[J]. 中国农业科学,2008,41(8):2485-2492.
[12] 陈渊,袁哲明,周玮,等. 基于地统计学与支持向量机的 QSAR 建模[J]. 物理化学学报,2009,25(8):1587-1592.
[13] 李哈滨,王政权,王庆成. 空间异质性定量研究理论与方法[J]. 应用生态学报,1998,9(6):651-657.
[14] LIEBHOLD A M,ROSSI R E,KEMP W P. Geostatistics and geographic information systems in applied insect ecology[J]. Annu Rev Entomo,1993,38:303-327.
[15] 王海燕,卢山. 非线性时间序列分析及其应用[M]. 北京:科学出版社,2006:139-142.
[16] CHANG C C,LIN C J. LIBSVM:A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(27):1-27.