基于遗传算法改进的 BP 神经网络房价预测分析

李春生,李霄野,张可佳
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:使用传统的 BP 神经网络进行预测容易发生收敛速度慢、预测精度低、陷入局部最优的可能。对此,阐述了 BP 神经网络的基本原理,介绍了遗传算法的实现过程,并根据遗传算法的全局搜索能力,优化调整了 BP 神经网络的初始权值和阈值,分别对传统 BP 神经网络和改进后的 GA-BP 神经网络建立了房价预测模型。选取了中国房价及其主要影响因素作为实验数据进行仿真训练,对比了模型的预测效果。实验结果表明,经过遗传算法改进的 BP 神经网络较传统 BP 神经网络具有预测精度高、收敛速度快的优点,同时避免了陷入局部最优的缺陷。

关键词:BP 神经网络;遗传算法;价格预测;误差分析
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2018)08-0144-04
doi:10.3969/j.issn.1673-629X.2018.08.030

Price Forecasting Analysis of BP Neural Network Based on Improved Genetic Algorithm

LI Chun-sheng, LI Xiao-ye, ZHANG Ke-jia
(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: Usingtraditional BP neural network for prediction is prone to be slow convergence, low prediction accuracy and easy to fall into local optimum. For this we describe the basic principles of BP neural network, introduce the implementation of genetic algorithm, and adjust the initial weights and thresholds of BP neural network according to the global search ability of genetic algorithm. Respectively, the traditional BP neural network and the improved GA-BP neural network are used to establish the housing price prediction model. Finally, the housing price and its main influencing factors are selected as the experimental data for simulation training, and the prediction effect of the model is compared. The experiment shows that the improved BP neural network has higher prediction accuracy and faster convergence speed than the traditional BP neural network, and avoids the defects of falling into the local optimum.

Key words: BP neural network; genetic algorithm; price forecasting; error analysis

0 引 言

近年来,国内房地产行业高速发展,房价的走势受到越来越多人的关注。研究房价的影响因素,并对未来房价进行预测,对于国家经济发展和改善民生具有重要的意义[1]。

国家每年都有相关的年鉴数据,如何从大量且无规律的房地产统计数据中挖掘出房价的走向,成为当今房地产行业的研究热点。文中提出了一种基于遗传算法改进的 BP 神经网络房价预测模型,以 2005-2012 年中国统计年鉴数据为基础,分析预测了 2013 年、2014 年、2015 年房价,并与相关真实数据进行比较,从而进行误差分析。

1 BP 神经网络基本原理

BP(back propagation)神经网络也称反向传播网络,是一种按误差逆传播算法训练的多层前馈网络[2],通常具有 3 层或 3 层以上结构。图 1 是一种简单的 3 层 BP 神经网络拓扑结构,包括输入层、输出层和隐含层,各层神经元与下层所有神经元连接,而同层各神经元之间无连接。图 1 中 x_1、x_2、x_3 为输入信号值,y_1、y_2、y_3 为输出信号值。

BP 神经网络的基本原理是采用梯度下降法,通过反向传播不断调整网络的权值和阈值,直到误差减小到最低[3]。其训练过程实质上是对各连接权值的动态调整,输入信号的正向传播与误差的反向传播过程循

收稿日期:2017-09-16 **修回日期**:2018-01-09 **网络出版时间**:2018-04-28
基金项目:黑龙江省自然科学基金面上项目(F2015020);省教育科研规划重点课题(GJB1215013)
作者简介:李春生(1960-),男,博士,教授,博导,研究方向为人工智能及其应用、模式识别与人工智能;李霄野(1993-),女,硕士研究生,通讯作者,研究方向为数据挖掘与人工智能。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20180427.1640.050.html

环进行,一直到输出的均方误差达到要求的标准。

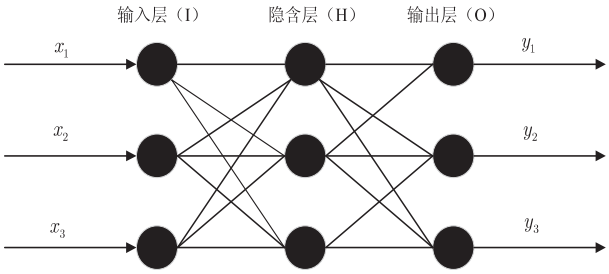

图 1　3 层 BP 神经网络拓扑结构

目标函数定义为:

$$E = 1/2\sum_{q=1}^{N}\sum_{k=1}^{m}(y_k^q - v_k^q)^2 \tag{1}$$

其中,y_k^q 为输出节点 k 在样本 q 作用时的期望输出;v_k^q 为输出节点 k 在样本 q 作用时的实际输出向量;N 为样本数据个数;m 为输出向量的维数。

BP 神经网络的训练过程如下:

(1)随机数初始化误差函数:假设网络结构为一个含有 n 个神经元的输入层,含有 1 个节点数的隐含层,含有 m 个节点数的输出层。输入层与隐含层之间的连接权值为 w_{ij},隐含层与输出层之间的连接权值为 w_{jk},隐含层各神经元的阈值为 $a = \{a_1, a_2, \cdots, a_l\}$,输出层各神经元的阈值为 $b = \{b_1, b_2, \cdots, b_m\}$。

(2)计算隐含层输出 h_j。

$$h_j = f(\sum_{i=1}^{n} w_{ij}x_i - a_j), j = 1, 2, \cdots, l \tag{2}$$

其中,f 为隐含层激励函数;x_i 为第 i 个输入节点变量。

(3)计算输出层函数 v_k。

$$v_k = \sum_{j=1}^{l} h_j w_{jk} - b_k, k = 1, 2, \cdots, m \tag{3}$$

(4)修正输出层连接权值。

$$w_{ij}(t+1) = w_{ij}(t) + \eta[(1 - mc)D(t) + mc * D(t-1)], i = 1, 2, \cdots, n \tag{4}$$

$$w_{jk}(t+1) = w_{jk}(t) + \eta[(1 - mc)D'(t) + mc * D'(t-1)] \tag{5}$$

其中,学习速率 $\eta > 0$,$D(t) = -\partial J/\partial w_{ij}(t)$,$D'(t) = -\partial J/\partial w_{jk}(t)$,$0 < mc < 0.9$。

动量因子 mc 的引入修正了神经元的权值,使其具有惯性和震荡能力,并根据反向传播法来产生新的权值变化,进而提高收敛速度[4]。

(5)修正隐藏层连接权值:根据网络实际输出 v_k 和期望输出 y_k 之间的误差更新 a_j,b_k。

$$a_j(t+1) = a_j(t) + \eta h_j(1 - h_j)\sum_{k=1}^{m} w_{jk}(y_k - v_k) \tag{6}$$

$$b_k(t+1) = b_k(t) + (y_k - v_k) \tag{7}$$

(6)计算全局误差,判断迭代是否结束,若未结束,返回步骤 2。

2　基于传统 BP 神经网络的房价预测模型

2.1　实验数据准备及数据预处理

实验数据的选择需要尽可能准确地反映房价的变化规律,在该研究中,选取影响中国房价的 7 个主要影响因素:商品房平均销售价格、住宅、居民消费水平、城镇单位就业人员平均工资、城镇居民可支配收入、国内生产总值 GDP 和城镇单位就业人员数。数据来源于《2016 年中国统计年鉴》中 2005-2015 年数据,具体如表 1 所示。

表 1　2005 年-2015 年中国房价及其相关影响因素数据

年份	商品房平均销售价格/(元/平方米)	商品房销售面积/万平方米	居民消费水平/元	城镇单位就业人员平均工资(元)	城镇居民可支配收入(元)	国内生产总值 GDP(亿元)	城镇单位就业人员数
2005	3 168	55 486.22	5 771	18 200	10 493.0	187 318.9	11 404.0
2006	3 367	61 857.07	6 416	20 856	11 759.5	219 438.5	11 713.2
2007	3 864	77 354.72	7 572	24 721	13 785.8	270 232.3	12 024.4
2008	3 800	65 969.83	8 707	28 898	15 780.8	319 515.5	12 192.5
2009	4 681	94 755.00	9 514	32 244	17 174.7	349 081.4	12 573.0
2010	5 032	104 764.65	10 919	36 539	19 109.4	413 030.3	13 051.5
2011	5 357	109 366.75	13 134	41 799	21 809.8	489 300.6	14 413.3
2012	5 791	111 303.65	14 699	46 769	24 564.7	540 367.4	15 236.4
2013	6 237	130 550.69	16 190	51 483	26 955.1	595 244.4	18 018.4
2014	6 324	120 648.54	17 778	56 360	29 381.0	643 974.0	18 277.8
2015	6 793	128 494.97	19 308	62 029	31 790.3	685 505.8	18 062.5

原始数据由于各项指标的数量级差别较大且量纲不同,为提高网络的训练效率,采用 Matlab 的归一化函数 premnmx,将实验数据的输入和输出数据归一化,把数据变换到[-1,1]的范围之间[5]。假设原始数据的输入样本为 i ,输出样本为 o ,用 premnmx 函数分别求出输入输出样本的最大值 max$i$ 和 max$o$,最小值 min$i$ 和 min$o$,利用式 8 和式 9 求出归一化处理后的输入样本 P_n 和输出样本 T_n :

$$P_n = \frac{2(i - \min i)}{\max i - \min i} - 1 \tag{8}$$

$$T_n = \frac{2(o - \min o)}{\max o - \min o} - 1 \tag{9}$$

在 BP 网络训练结束后,对于得到的归一化数据,需要用 postmnmx 函数对其进行反归一化处理,还原成正常值[6]。

2.2 建立训练模型

选取归一化处理后的中国房价样本数据,前两年数据作为输入向量,两年后的第一年数据作为目标输出,如用 2005、2006 年数据预测 2007 年房价,用 2006、2007 年数据预测 2008 年房价,以此类推,共得出 9 组数据。将 2005 - 2012 年数据作为训练样本, 2013 - 2015 年数据作为测试验证样本。选定动量因子 mc = 0.8,最大训练步数 8 000 次,误差设定值 0.001,建立仿真模型。

$$H_n = \sqrt{I_n + O_t} + \alpha \tag{10}$$

其中, H_n 为隐含层节点数; I_n 为输入层节点数; O_t 为输出层节点数; α 为 1 ~ 10 之间的整数。

根据前面分析,输入层神经元个数为 9,输出层神经元个数为 1,经过计算后得出隐含层神经元个数 H_n 范围大概在 4 ~ 13 之间;而隐含层神经元个数的数值过大或过小都会导致网络误差震荡或收敛时间长。综合这两个因素,最终选取隐含层神经元个数 H_n 为 8。

BP 神经网络的各隐含层节点的激活函数选取 Sigmoid 函数,输出层节点的激活函数采用对 S 型 Sigmoid 函数[7]。

2.3 BP 网络的模型训练和结果

实验采用 Matlab2011b 中的神经网络工具箱实现 BP 神经网络模型的构建、训练和仿真,将 2005 - 2012 年各房价指标数据输入到设计好的 BP 神经网络,得到网络的输出值,与实际值进行比较,再不断调整权值和阈值,直到计算的误差值达到要求的范围。再将 2013 - 2015 年房价数据作为检验样本输入,以此判断实验结果,对结果进行还原和分析,得到实际值。预测值与真实值的对比如表 2 所示,可知训练的最低准确度为 96.89%。

表 2 BP 神经网络模型的房价预测结果

年份	实际房价 /元	预测房价 /元	准确度 /%
2013	6 237	6 365.37	97.94
2014	6 324	6 227.21	98.47
2015	6 793	6 581.56	96.89

3 遗传算法改进的 BP 神经网络模型

遗传算法(GA)是受达尔文进化论的启发,借鉴生物进化过程而提出的一种启发式搜索算法。它将要解决的问题模拟成一个生物进化的过程,通过种群之间进行选择、交叉和变异等操作,逐步淘汰适应度函数值低的个体,增加适应度函数值高的个体[8]。经过多代循环,最终产生出符合条件的个体。BP 神经网络在逼近预测值过程中耗时多,导致网络收敛速度慢,且容易陷入局部最优。而遗传算法则具有高容错性、并行、全局择优的特点[9]。因此,文中利用遗传算法来优化 BP 神经网络大的初始权值和阈值,形成 GA-BP 神经网络,提高 BP 神经网络的收敛速度,降低 BP 算法陷入局部最优的可能性。

为了对比 BP 神经网络与 GA-BP 神经网络,在训练 GA-BP 神经网络时,参数选取与之前的 BP 神经网络一致。

3.1 遗传算法改进 BP 神经网络的步骤

(1)种群初始化。

将网络中所有的权值和阈值进行实数编码,每个个体作为一组染色体[10],染色体形式为:

$$w_{11}, w_{12}, \cdots, w_{ij}, a_1, a_2, \cdots, a_l, w_{11}, w_{12}, \cdots, w_{jk}, b_1, b_2, \cdots, b_m$$

其中, w_{ij} 为输入层与隐含层的连接权值; $a = \{a_1, a_2, \cdots, a_l\}$ 为隐含层阈值; w_{jk} 为隐含层与输出层的连接权值; $b = \{b_1, b_2, \cdots, b_m\}$ 为输出层阈值。

遗传算法的全局搜索性能很大程度上受种群数量的影响,种群的数量要根据具体问题来选取[11]。因此,此次实验初始种群的规模为 100。

(2)选择适应度函数。

BP 神经网络中的误差绝对值和越小越好,而在遗传算法中,适应度值越大越好。因此,以 BP 神经网络目标函数的倒数作为适应度函数,即:

$$F(x) = \left(\sum_{q=1}^{N} \sqrt{\sum_{k=1}^{m} (y_k^q - v_k^q)^2} \right)^{-1} \tag{11}$$

(3)基因选择。

采用轮盘赌法对种群中的个体进行选择操作,选择适应度高的个体遗传到下一代[12],每个个体 x 被选择的概率 p_x 为:

$$p_x = \frac{F_x}{\sum_{x=1}^{k} F_x} \tag{12}$$

其中,k 为种群个体的数目;$F(x)$ 为个体 x 的适应度值。

(4)交叉与变异操作。

交叉运算是遗传算法中产生新个体的主要操作过程,目的是通过使用交叉算子从全局的角度改善个体编码结构[13]。变异操作是对群体中的个体串的某些基因座上的基因值作变动,可以产生新的个体,使遗传算法具有局部的随机搜索能力[14]。

(5)循环操作。

当个体的适应度达到给定的阈值,或者个体和群体的适应度不再上升时,算法的迭代过程收敛、算法结束。否则,返回到第 2 步执行循环。将遗传算法优化后的连接权值和阈值作为 BP 神经网络的初始权值和阈值,进行 GA-BP 神经网络训练,直到满足误差要求或达到最大训练次数为止[15]。

3.2　改进遗传算法的 BP 神经网络模型的训练结果

将改进的 GA-BP 神经网络模型进行训练,对 2013-2015 年 3 组测试验证样本进行预测。利用仿真实验得出的预测结果与实际房价进行比较,如表 3 所示。对比得出,GA-BP 神经网络模型对 2013-2015 年房价预测的最低准确度为 99.20%。

表 3　遗传算法改进 BP 神经网络模型的房价预测结果

年份	实际房价/元	预测房价/元	准确度/%
2013	6 237	6 285.20	99.32
2014	6 324	6 301.56	99.65
2015	6 793	6 738.65	99.20

4　实验结果对比

经过 BP 神经网络模型与 GA-BP 神经网络模型的实验结果对比,可以得出各房价预测模型的误差统计,如表 4 所示。BP 神经网络和 GA-BP 神经网络的误差曲线对比如图 2 所示。

表 4　BP 神经网络与 GA-BP 神经网络的房价预测误差对比　　%

模型	最小相对误差	最大相对误差	平均相对误差
BP 神经网络	1.53	3.11	2.32
GA-BP 神经网络	0.35	0.8	0.575

由表 4 得出,传统 BP 神经网络的房价预测模型的误差远高于 GA-BP 神经模型;从图 2 可以看出,当最大训练步数为 8 000 次,目标误差设定为 0.001 时,GA-BP 神经网络经过的训练次数明显少于 BP 神经网络,且误差曲线趋势较为平缓,更加贴近目标曲线,收敛速度明显加快。由此可得出,遗传算法改进后的 GA-BP 神经网络的预测值更接近于真实值。

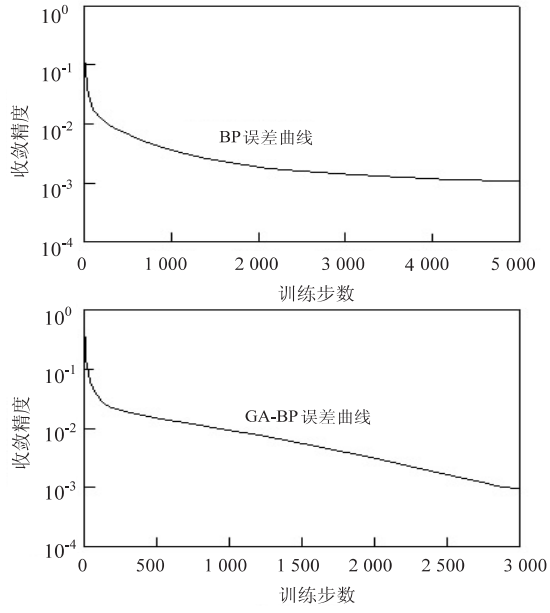

图 2　BP 神经网络模型(上)与 GA-BP 神经网络模型(下)误差曲线对比

5　结束语

利用遗传算法的全局搜索能力找出 BP 神经网络的最优初始权值和阈值,进而使 BP 神经网络具有更快的收敛速度和更高的精度。采用 2005-2012 年中国房价及其主要影响因素作为实验数据,分别用 BP 神经网络和 GA-BP 神经网络对 2013-2015 年房价进行了预测。结果表明,GA-BP 神经网络具有预测精度高、收敛速度快的特点,同时避免了陷入局部最优的缺陷,可以作为房价预测的一种可靠的方法,同时也可以尝试作为其他类型数据的预测方法。

参考文献:

[1] 吴晓彬,沈陈华,浦晓天,等. 基于 BP 神经网络模型的南京市住房发展水平研究[J]. 南京晓庄学院学报,2007,23(3):77-81.

[2] 王雅楠,孟晓景. 基于动量 BP 算法的神经网络房价预测研究[J]. 软件导刊,2015,14(2):59-61.

[3] 周学君,陈文秀. 基于人工神经网络 BP 算法的黄冈市房价预测[J]. 黄冈师范学院学报,2014,34(3):13-15.

[4] ZHANG Sirui, WANG Botao, LI Xueen, et al. Research and application of improved gas concentration prediction model based on grey theory and BP neural network in digital mine

(下转第 151 页)

参考文献:

[1] 陈世敏. 大数据分析与高速数据更新[J]. 计算机研究与发展,2015,52(2):333-342.

[2] 杨宏波. 旅游信息系统中的防冲突任务调度模型仿真[J]. 计算机仿真,2015,32(6):447-451.

[3] 李　吟. 基于接口契约的有状态 Web 服务用例集生成[J]. 计算机研究与发展,2017,54(3):609-622.

[4] 武永斌,李国清,卢小平. 基于 NewMap Server 的旅游信息服务系统开发与实现[J]. 地理信息世界,2016,23(5):104-108.

[5] 王星捷,刘华春,李春花. 基于多元平台洪灾报警系统设计与实现[J]. 计算机技术与发展,2017,27(4):196-199.

[6] 王星捷,杨　森. ArcGIS Server 分布式技术实现与优化[J]. 计算机工程与设计,2012,33(9):3432-3436.

[7] 王星捷. 基于 MapGIS 三维数字城市的实现[J]. 计算机技术与发展,2016,26(12):96-98.

[8] 黄书力,胡大裟,蒋玉明. 经过指定的中间节点集的最短路径算法[J]. 计算机工程与应用,2015,51(11):41-46.

[9] 姬鹏飞,李远刚,卢盛祺,等. 基于语义 Web 的旅游路线个性化定制系统[J]. 计算机工程,2016,42(10):308-317.

[10] 孙祖汉,李　莹,罗智凌,等. 可视化 REST 服务组合框架的设计与实现[J]. 小型微型计算机系统,2017,38(1):10-14.

[11] 陈泰生,陈梦琳,孙敬杰,等. 一种 GIS 共享图形软件符号的方法[J]. 测绘科学,2016,41(8):116-120.

[12] JAYAKUMAR K, MALARVANNAN S. Assessment of shoreline changes over the Northern Tamil Nadu Coast, South India using WebGIS techniques[J]. Journal of Coastal Conservation,2016,20(6):1-11.

[13] LI Rui,FAN Jiapei,JIANG Jie,et al. Spatiotemporal correlation in WebGIS group-user intensive access patterns[J]. International Journal of Geographical Information Science,2017,31(1):36-55.

[14] WANG Jiechen,NI Haochen,RUI Yikang,et al. A WebGIS-based teaching assistant system for geography field practice (TASGFP)[J]. British Journal of Educational Technology,2016,47(2):279-293.

[15] DHAMANIYA A, SONU M, KRISHNANUNNI M, et al. Development of web based road accident data management system in GIS environment: a case study[J]. Journal of the Indian Society of Remote Sensing,2016,44(5):1-8.

(上接第 147 页)

[J]. Procedia CIRP,2016,56:471-475.

[5] WEN Yuechun, TAN Lina, WU Hailong. Inflation forecast ba-sed on BP neural network model[J]. Advanced Materials Research,2014,989-994:5536-5539.

[6] 王德明,王　莉,张广明. 基于遗传 BP 神经网络的短期风速预测模型[J]. 浙江大学学报:工学版,2012,46(5):837-841.

[7] 高玉明,张仁津. 基于遗传算法和 BP 神经网络的房价预测分析[J]. 计算机工程,2014,40(4):187-191.

[8] 刘奕君,赵　强,郝文利. 基于遗传算法优化 BP 神经网络的瓦斯浓度预测研究[J]. 矿业安全与环保,2015,42(2):56-60.

[9] 刘智斌,曾晓勤,刘惠义,等. 基于 BP 神经网络的双层启发式强化学习方法[J]. 计算机研究与发展,2015,52(3):579-587.

[10] 尹光志,李铭辉,李文璞,等. 基于改进 BP 神经网络的煤体瓦斯渗透率预测模型[J]. 煤炭学报,2013,38(7):1179-1184.

[11] 刘春艳,凌建春,寇林元,等. GA-BP 神经网络与 BP 神经网络性能比较[J]. 中国卫生统计,2013,30(2):173-176.

[12] SUN Weiwei, YAO Yunfei, WANG Chunsheng, et al. BP network optimization based on improved genetic algorithm[J]. Advanced Materials Research, 2012, 532-533: 1757-1763.

[13] KE Gang,HONG Yinghan. The research of network intrusion detection technology based on genetic algorithm and BP neural network[J]. Applied Mechanics and Materials,2014,599-601:726-730.

[14] 付娅丽. 遗传算法在指纹识别特征匹配中的应用[D]. 北京:北京邮电大学,2006.

[15] 孙玲芳,周加波,林伟健,等. 基于 BP 神经网络和遗传算法的网络舆情危机预警研究[J]. 情报杂志,2014,33(11):18-24.

稿　件　订　正

在本刊 2018 年 6 月份增刊 126-130 页上刊登的曹良坤,林锦屏,罗裕梅,郭艳琴,王林茂的论文《基于 AHP 的避寒旅游地交通适宜度评价方法研究》中,由于编辑人员的失误,漏掉了基金项目和通讯作者简介,现在对论文做补充订正。

基金项目:国家自然科学基金资助项目(41561031)

通讯作者:林锦屏(1963-),女,福建省福州人,副研究员,硕士生导师,研究方向为区域旅游与人文地理。