

附件 2：李英冰、陈雨劲、欧阳茜，基于 BP 神经网络的武汉市二手房估价模型研究，数字制造科学（录用期刊，文章编号：1672-3236（2017）01-02-0066-05）



## 基于BP神经网络的武汉市二手房估价模型研究

李英冰,陈雨劲,欧阳茜

(武汉大学测绘学院,湖北 武汉 430069)

**摘 要:**针对传统的估价方法主要是基于市场交易的简易计算,没有着重考虑房地产价格受区位特征、建筑特征和邻里特征的影响,且带有较强的主观性等问题,采用网络爬虫技术获取到武汉市大量的二手房信息,并剔除异常值,然后将房地产价格的影响因素进行量化,得到模型的样本集。以特征价格模型为基础,建立了基于BP神经网络的房地产估价模型。将样本集分为训练集、验证集和测试集,实验使用训练集训练模型的参数,用验证集计算训练过程中的准确率以控制终止迭代,并用测试样本进行检验,发现该模型有较好的估价效果。

**关键词:**BP神经网络;特征价格模型;房地产估价;数据挖掘

**中图分类号:**F224.9

DOI:10.3963/j.issn.1672-3236.2017.01-02.013

我国房地产估价伴随土地改革和商品房的出现而诞生于20世纪80年代末,房地产价格受制于政府经济计划和地域的限制,具有很强的地域特性和偶然性。不同规模、不同性质的住宅往往差别很大,难以简单地套用线性化公式模型来估价,数学模型的建立也面临很多困难。这就要求评估人员对评估对象作出更加准确的评估,科学地进行房地产评估是一个十分重要的课题。

一宗房地产的价格通常采用市场比较法、成本法和收益法来进行评估<sup>[1]</sup>。这些方法是一些基于市场交易的简化计算方法,没有考虑影响商品住宅价格的因素。特征价格模型将二手房价格分解为各个特征价格,以各个特征对房价的影响为根据,从而实现房地产估价<sup>[2]</sup>。对于同期房地产价格而言,房地产价格受区位特征、建筑特征和邻里特征三大特征变量影响<sup>[3]</sup>。

房地产价格评估的核心要素是已有数据,大数据不仅为房地产估价提供了海量的数据,也影响着房产估价的方式<sup>[4]</sup>。大数据时代之前关注点在于分析数据之间的因果关系,现在则更加注重海量数据之间的相关关系。使用大量的量化数据用于训练评估模型,可以有效降低人为主观因素的影响,提高价格预测的准确性和客观性。

近年来出现较多基于数据的房地产估价方法

的研究,通过特征数据内部隐藏的相关关系进行估价。基于区间数的灰色模糊房地产估价方法是对市场比较法的一般性改进,但估价准确性依赖于类似的交易数据<sup>[5]</sup>。多元线性回归方法将价格影响因素作为自变量,将房产价格作为因变量,通过对数据的分析研究,寻找变量间的依赖关系,建立基于多元线性回归的房地产价格模型,但线性模型的拟合结果难以满足要求<sup>[6]</sup>。近年来机器学习的方法也被应用到房地产估价领域,运用支持向量回归构建二手房批量评估模型,可得到较高的预测准确率<sup>[7]</sup>。

机器学习重在寻找数据中的模式,并使用这些模式来做出预测。在机器学习领域,神经网络是一种有效用于对函数进行估计和近似计算的模型。鉴于此,笔者以武汉市二手房普通住宅为例,结合大数据的思想,提取二手房的信息并进行数据预处理,然后利用BP神经网络对数据进行建模,并可通过该模型进行武汉市二手房价格预测。

### 1 模型构建

#### 1.1 BP神经网络模型

人工神经网络(artificial neural network, ANN)是一种模仿生物神经网络的结构和功能的计算模型,由大量的人工神经元联结进行计算,用

收稿日期:2017-03-20

作者简介:李英冰(1972-),男,湖北十堰人,武汉大学测绘学院副教授,博士,硕士生导师。

于对函数进行估计和近似,是一种自适应系统。神经网络拥有一种普遍性,单个隐含层的神经网络可用于拟合任意函数<sup>[8]</sup>。

BP(back propagation)神经网络是一种多层前馈网络,是目前应用最为广泛的神经网络模型之一。通过对网络中所有权重计算损失函数的梯度,反馈给最优化方法,用于更新权重以最小化损失函数,迭代进行以上过程直到网络对输入的响应达到满意的预定目标范围为止。单隐层神经网络如图1所示。



图1 单隐层神经网络

### 1.2 模型变量的选取与量化

在建立二手房估价模型之前,需要选取房地产价格对应的解释变量。结合特征价格模型,选取面积、楼层、房间数、朝向、建成年份、装修程度、医院、商场、公园、学校、公交线路、地铁、商圈和小区容积率这14个解释变量进行研究,并采用了一定的量化方法对这些变量进行量化整合,结果如表1所示。

表1 二手房价格解释变量的选取与量化

特征分类	特征变量	内容
区位特征	公交	小区周围1 000 m 经过的公交线路数
	地铁	如果5 000 m 内无地铁站,则赋值为5,否则等于与最近的地铁站的距离/1 000
	商圈	如果5 000 m 内无商圈,则赋值为5,否则等于与最近商圈的距离/1 000
	面积	一套二手房的总建筑面积(平方米)
建筑特征	楼层	所在楼层的层数
	房间数	客厅总数
	朝向	南北朝向(1)、其他(0)
	建成年份	楼盘竣工年份(年)
邻里特征	装修程度	毛坯(1)、简装修(2)、中(3)、精(4)、豪华(5)
	容积率	小区容积率
	学校	1 000 m 内学校数目
	医院	2 000 m 内医院数目
特征	商场	2 000 m 内商场数目
	公园	1 000 m 内公园数目

### 1.3 武汉市二手房估价模型建立

#### 1.3.1 BP神经网络模型结构设计

选取BP神经网络作为建模工具,采用单隐层的BP网络结构。输入层的神经元数等于房地产价格影响因素的个数,即14个;输出的结果为预测的房地产价格,因此输出层神经元数为1;隐含层的神经元个数由Hornik提出的公式: $N = [\sqrt{2n+m}, 2n+m]$ 确定,其中 $N$ 为神经网络隐含层的神经元个数, $n$ 为输入层节点个数, $m$ 为输出层节点个数,经过试验发现本实验中隐含层的神经元个数为25时,神经网络预测效果较好。实验选取LM算法(Levenberg-Marquardt method)作为网络的主要训练方法<sup>[9]</sup>,实验设计的BP神经网络模型结构如图2所示。

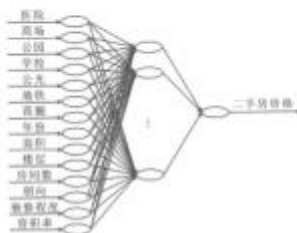


图2 实验设计的BP神经网络模型结构

#### 1.3.2 BP神经网络进行房地产估价的过程

BP神经网络包括信号的前向传播和误差的反向传播两个过程。输入信号从输入层经隐含层逐层处理,转向输出层,产生输出信号,若实际输出与期望输出不符,则转入误差的反向传播,将误差信号沿原来的连接通路返回,通过修改各神经元的权值,使误差沿梯度方向下降,经过反复学习训练,确定与最小误差相对应的网络参数,即可停止训练。

在进行房地产估价建模时,将可用二手房数据样本分为训练集、验证集和测试集。在训练网络时,通过训练集来训练神经网络模型中的参数,训练好的模型对训练集样本的符合度称为模型的可塑性,对测试集样本的符合度称为模型的泛化能力。性能好的模型需要同时具备良好的可塑性和泛化能力。笔者采用提前终止法来防止模型出现“过拟合”,增加了验证集,在每个迭代期的最后计算在验证集上的准确率,一旦准确率达到饱和就停止训练,即得到武汉市二手房估价模型。

## 2 实证研究

### 2.1 实验数据预处理

#### 2.1.1 异常值剔除

异常值是指样本中的个别值,其数值明显偏离其余的观测值。在本实验中采取的是网络爬虫获取的数据作为样本,存在一些大幅偏离正常值的数据,这些数据的存在会影响到模型的效果,需要对异常值进行剔除。

由爬虫得到的武汉市各小区的二手房信息覆盖面广,数据量大,但由于各小区之间差异较大,数据的差别也比较大,难以统一进行异常值的剔除。而对于单个小区的数据,房地产的交易价格的差异相对较小,易于通过分析剔除异常值。实验中,将样本分小区进行异常值的探测,对于单个小区内的数据,先计算小区内交易价格的均值和中误差。经过分析发现,受房地产自身因素的影响,同一小区内成交价也会存在较大差异,但绝大部分差异都在2倍中误差之内,有少量的交易价格与均值的差值远大于2倍中误差,将这些数据判断为异常值。以均值加减两倍中误差的范围作为阈值,剔除各小区内数据的异常值。

#### 2.1.2 输入数据标准化

对于神经网络的输入样本,各因素之间在数量上有数量级的差异,进行预处理后可使样本数据的数值在同样的数量级。假设输入数据以  $X = (x_1, x_2, x_3, \dots, x_n)$  表示,输出数据以  $Y = (y)$  表示,其中  $x_1, x_2, \dots, x_n$  表示房地产价格的影响因素,  $y$  表示房地产价格。本实验采用标准化方法对输入数据进行预处理,标准化之后样本均值为0,方差为1。标准化变换式为:

$$\bar{X}_i = \frac{X_i - \bar{X}}{\sqrt{\text{var}(X_i)}} \quad (1)$$

式中,  $\bar{X}_i$ 、 $\sqrt{\text{var}(X_i)}$  为样本的均值和标准差。

### 2.2 实验结果与分析

#### 2.2.1 模型评价指标

(1) 可决系数 ( $R^2$ )。可决系数作为综合度量回归模型对样本观测值拟合优度的度量指标。可决系数越大,模型拟合优度越好。其计算公式如下:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \quad (2)$$

式中:  $ESS$  为回归平方和 (Explain Sum of Squares);  $TSS$  为总离差平方和 (Total Sum of Squares);  $RSS$  为残差平方和 (Residual Sum of Squares), 其计算公式如下:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (4)$$

式中:  $y_i$  为每个样本的实际价格;  $\bar{y}$  为样本均价;  $\hat{y}_i$  为每个样本的价格估计值。

模型的拟合程度还可以用相关系数  $R$  表示。可决系数为相关系数的平方。

(2) 平均绝对百分比误差 (MAPE)。平均绝对百分比误差是指参数估计值与参数真值的差值与参数真值之比的期望值,其计算公式如下:

$$MAPE = \frac{1}{T} \sum_{i=1}^T \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (5)$$

式中:  $T$  为样本数目;  $y_i$  为每个样本的实际价格;  $\hat{y}_i$  为每个样本的价格估计值。

#### 2.2.2 实验结果

经过139次迭代,达到目标误差,神经网络模型训练成功,计算结果如表2所示,图3分别表示模型在训练集、验证集、测试集和全部数据集上的回归情况,其中  $R$  表示回归值 (相关系数),图4表示误差分布情况。

表2 BP神经网络模型训练结果

样本	样本数	MAPE/%	$R^2$	$R$
训练集	30 873	6.97	0.936 1	0.967 5
验证集	6 615	7.18	0.933 8	0.966 3
测试集	6 615	7.02	0.934 3	0.966 6

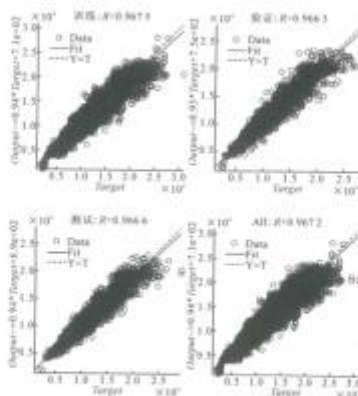


图3 模型回归情况界面图

从模型输出结果来看,在测试集上的预测价格与真实交易价格的平均绝对百分比误差为

7%,模型整体的相关系数为0.966 6,可见BP神经网络模型的预测效果较好。

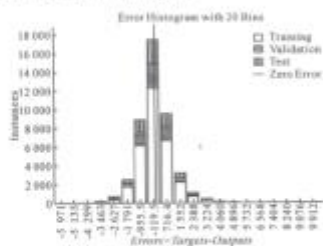


图4 误差分布直方图

### 3 讨论

在同一个小区内,除了本文考虑的楼层、房龄数和房屋朝向等因素对于房地产价格的影响,还有一些难以放入模型的参数,如楼座的具体位置与小区环境的关系。同一小区内楼座位置不同,其景观和噪声、灰尘污染也不同,拥有湖景、山景的楼座,价格会更高,噪声污染、灰尘污染比较大的楼座,价格就会偏低。还有在一些面积较大的小区,不同楼座距离小区门口或者附近公交站或地铁站的距离差别会比较大,房源的价格也会有差异。网络爬取的数据,数据来源广泛,装修程度的量级、实际交易价格等受售房者主观性影响较大。这些因素也导致了房地产估价的精度受限。

实验对样本的异常值进行剔除,对比剔除异常值之前,将测试集平均绝对百分比误差由9.98%降至7.02%,相关系数由0.921 0提高到0.966 6。样本各部分的平均绝对百分比误差、可决系数和相关系数均有较大提高。

廖格等也应用特征价格模型,选取多元线性模型和对数模型的方法对武汉市二手房价格进行研究<sup>[10]</sup>,多元线性回归实验结果的可决系数为0.86,多元对数模型实验的可决系数为0.87,本文采用BP神经网络模型的可决系数为0.93,模型拟合情况优于多元线性模型和多元对数模型。

### 4 结论

基于特征价格模型,结合武汉市二手房数据,采用BP神经网络的方法,构建了武汉市二手房价格预测模型,经过模型的可决系数、相关系数和平均绝对百分比误差等验证,BP神经网络模型对数据模型的拟合效果较好。

BP神经网络建模的方法是数据驱动的,对于数据的分析需要进一步研究。网络爬虫获取的数据具有数据量大的特点,但数据的质量良莠不齐,对获取的数据进行筛选得到质量较高的数据对于模型的构建十分重要。以后的研究重点可放在如何对爬虫获取的多维数据进行质量检验,并设计出高效的异常点识别算法。

### 参考文献:

- [1] 王彦伟. 房地产估价理论与应用[M]. 北京:清华大学出版社,2014.
- [2] Caut A. Hedonic Price Indexes with Automotive Examples[J]. In: The Dynamics of Automobile Demand, 1939(1):99-117.
- [3] Butler R V. The Specification of Hedonic Indexes for Urban Housing[J]. Land Economics, 1982,58:94-108.
- [4] 周丹,郭化林. 大数据时代对商业地产估价的影响研究[J]. 商场现代化,2014,28:254-255.
- [5] 吴红华,赖华勇. 房地产估价的区间数灰色模糊法[J]. 湖南大学学报(自然科学版),2012,39:27-30.
- [6] 孙小磊. 基于多元线性回归分析法的房地产价格评估[J]. 商业经济研究,2014,36:133-134.
- [7] 宋祖杰. 基于支持向量回归的二手房批量评估模型应用研究[D]. 重庆:重庆大学,2016.
- [8] Hornik K, Stinchcombe M, White H. Multilayer Feedforward Networks are Universal Approximators[J]. Neural Networks, 1989,2(5):359-366.
- [9] Gavin H P. The Levenberg - Marquardt Method for Nonlinear Least Squares Curve-fitting Problems[D]. Department of Civil and Environmental Engineering, Duke University,2013.
- [10] 廖格,李英冰,袁菲. 基于多元回归法的武汉市二手房价格影响因素研究[J]. 城市勘测,2017(1):33-38.

### Appraisal Model based on BP Neural Network Method for Second-hand House in Wuhan

LI Yingbù, CHEN Yujūn, OUYANG Xi

**Abstract:** Real estate prices are subject to locational, architectural and neighbourhood characteristics. Conventional appraising methods are limited by subjectivity, and simplified calculations upon market transactions. By removing outliers and quantifying the second-hand house data obtained by web crawlers, we obtained the sample set for the model in this study. Based on hedonic price model, a BP neural network method for real estate appraisal model has been built up. The samples were grouped by training set, validation set and testing set. The training set was used for training parameters in our model, while the validation set was for calculating the accuracy during the training stage and controlling the end of iterations. According to the results, it proves that the model introduced in this paper is applicable and reliable.

**Key words:** BP neural network; hedonic price model; real estate appraisal; data mining

LI Yingbù: Associate Professor; School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China.

[编辑:周廷美]

(上接第55页)

### Research on Abnormal Signal Processing of Magnetic Grid based on Improved Incremental Mean Method

LU Jun, XIAO Jun, YANG Zheng, HU Tianxi, ZHANG Xing, JIANG Ding

**Abstract:** In the practical working condition of the absolute magnetic grating displacement sensor, signal fault due to poor stability of the magnetic grid usually occurs. The abnormal value of the displacement signal returned by the magnetic grid occurs frequently, affecting the acquisition of real displacement data to a specific extent. To improve the reliability and provide accurate displacement data, this paper focuses on the principle and abnormal characteristics of magnetic grid signal. Authors develop an algorithm improvement method for the mean increment and the magnetic grid displacement signal features using the displacement signal of outlier mining method based on distance model. The experiment and practical application show that the developed method is effective in eliminating the influence of disturbance on the displacement return value and improving of the displacement accuracy.

**Key words:** the absolute magnetic grid displacement sensor; outlier; distance model; improved incremental mean method

LU Jun: Postgraduate; School of Mechanical and Electric Engineering, WUT, Wuhan 430070, China.

[编辑:周廷美]