

# MLP 模型在房地产批量评估中的应用研究

■ 仇晓洁 罗荣芸

(河北经贸大学财政税务学院, 河北石家庄 050061)

**【摘要】**第十三届全国人大常委会第三十一次会议通过《授权国务院在部分地区开展房地产税改革试点工作的决定(草案)》,这意味着房地产税改革进入实质性推进阶段。在此背景下,以批量评估为基础的房地产价值评估为房地产税的开征提供了技术支撑。本文运用人工神经网络中多层感知机(简称MLP)模型,以石家庄市长安区二手住宅为样本,进行实证研究,并将其结果与特征价格模型所得结果进行对比;同时,对裕华区二手住宅做以同样的实证分析。最终得出:相比经典特征价格模型,运用多层感知机模型得出的评估结果与真实交易值的误差更小,匹配度更优。

**【关键词】**批量评估 人工神经网络 多层感知机模型 特征价格模型

**【中图分类号】**F293

**【文献标识码】**A

**【文章编号】**1007-0265(2022)01-0024-07

## 一、引言

2021年3月,十三届全国人大四次会议通过的“十四五”规划和2035年远景目标纲要,明确将“推进房地产税立法”纳入工作重点<sup>[1]</sup>;5月,财政部、全国人大常委会预算工委、住房和城乡建设部、国家税务总局召开房地产税改革试点工作座谈会,听取部分城市及部分专家学者对房地产税改革试点工作的意见;10月,第十三届全国人大常委会第三十一次会议通过《授权国务院在部分地区开展房地产税改革试点工作的决定(草案)》;房地产税的开征渐行渐近。为满足房地产税税基评估的需求,需不断提升房地产税税基评估的批量评估技术。基于此,本文试图借助机器学习以及大数据资源,引入新的房地产批量评估技术——人工神经网络的多层

感知机,探索二手住宅的批量评估模型,以期为预期的房地产税征收提供计税依据。

## 二、房地产批量评估方法综述

房地产批量评估,就是指评估机构和人员在给定时间用标准化的方法,引用共同数据,进行测试等方法对大批量房地产进行评估的过程<sup>[2]</sup>。

通过梳理国内外文献发现,房地产批量评估中应用较早、且较为常用的方法是特征价格模型,常用形式是线性回归,它是基于特征价格理论构建的模型。该理论认为房价是由多种特征带给人们的效用决定的。国内外学者皆尝试利用此模型对房产税税基进行批量评估。其中,国外研究多集中于特征价格模型本身的优化改进,如用加法,乘法,混合模型回归方法(Connor, 2002)<sup>[3]</sup>,或是模型与

〔基金项目〕国家社科基金项目:SLRC2017008;河北省百名创新人才项目:15CJL058C。

〔作者简介〕仇晓洁,女,河北经贸大学财政税务学院,教授,硕士生导师,博士,研究方向为:资产评估、社会保障、财政理论与政策;罗荣芸:河北经贸大学财政税务学院,资产评估专业硕士,研究方向为资产评估。

传统市场法的效果对比 (John D. Benjamin, Randall S. Guttery, C. F. Sirmans, 2004)<sup>[4]</sup>。而国内研究主要针对变量关系、模型本身改进等方面展开研究,如用弹性分析影响因素间关系 (孙礼圣、邓宇, 2018)<sup>[6]</sup>, 利用空间计量优化特征价格模型 (金杰, 罗婷婷, 2021)<sup>[7]</sup>。

但伴随大数据时代的到来, 机器学习开始逐步被应用于各类领域中, 包括房地产批量评估。所谓机器学习, 即使用计算机模拟或实现人类学习活动的科学, 通过对数据或经验的学习实现自身改进。目前, 运用于房地产批量评估中的机器学习的算法, 主要有随机森林、支持向量机和人工神经网络等算法。其中, 随机森林是一个包含多个决策树的分类器, 其取多棵决策树组合在一起取预测值的平均值为结果, 国外学者较早尝试此方法, 通过构建随机森林模型对房地产价格进行评估 (EA Antipov, EB Pokryshevskaya, 2012)<sup>[8]</sup>。支持向量机算法是利用有限的样本信息进行模型训练。在房产税税基批量评估中, 支持向量机算法通常作为对比的方法出现 (司玺同, 2019)<sup>[9]</sup>。人工神经网络 (Artificial Neural Network, ANN) 是目前机器学习的研究热点, 具有自学习功能, 自适应性强。它通过模仿人类大脑神经元系统相互协作完成任务的过程, 引入激活函数刺激神经元, 将信息传递到下一层, 变成非线性的信息处理系统。人工神经网络主要包括多层感知机、BP 神经网络、卷积神经网络及递归神经网络等。而在房地产批量评估中应用较为广泛的人工神经网络模型是 BP 神经网络模型, 如收益性房产评估 (赵愈等, 2021)<sup>[10]</sup>, 城市二手住宅评估 (陈诗沁等, 2020)<sup>[11]</sup>。

与 BP 神经网络不同的是, 多层感知机是一种前馈神经网络。后者将输入的多数据集映射到单一的输出数据集, 克服了单层感知机不能对线性学习不可分数据进行识别的缺点。与 BP 神经网络相比, 多层感知机的运行和训练效率更快, 更多地是偏向一种结构, 而非算法。作为基础的神经网络, 后期可以对多层感知机添加多种算法不断优化, 可操作空间更大。但利用多层感知机进行批量评估的研究较少。

综上所述, 鉴于多层感知机的优势, 在此运用

多层感知机模型, 对石家庄市长安区二手住宅进行批量评估, 将特征价格模型作为参照模型, 对比两个模型的评估结果, 以验证多层感知机模型在批量评估中的有效性, 为税基批量评估提供新的思路。

另外, 为确保结论的准确性, 将此模型运用于石家庄市裕华区二手住宅的批量评估做进一步验证。

### 三、多层感知机 (MLP) 模型概述

多层感知机 (Multilayer Perceptron, 简称 MLP) 模型也称为多层前馈神经网络模型, 是人工神经网络模型的一种。所谓人工神经网络是一种可用于处理具有多个节点和多个输出点的实际问题的网络结构, 除了输入输出层, 它中间可以有多个隐层。多层感知机也具有同样的结构, 这一结构使其具有出色的非线性匹配和泛化能力, 如图 1<sup>[12]</sup>:

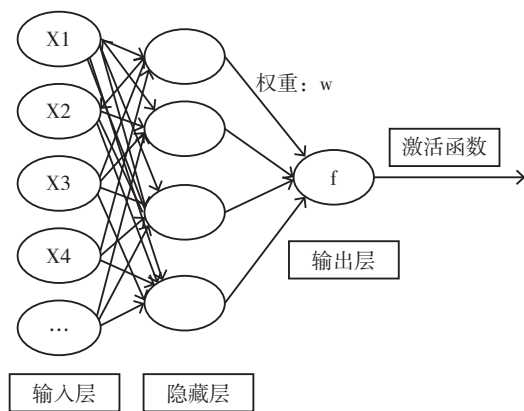


图 1 多层感知机结构

从图 1 可看出, 多层感知机层与层之间是全连接的。多层感知机最底层是输入层, 中间是隐藏层, 最后是输出层, 多层感知机是一种前向的神经网络, 它的输入是一组向量, 输出为另一组向量。多层感知机由输入层 (Input Layer)、隐藏层 (Hidden Layer) 和输出层 (Output Layer) 构成。其中, 隐藏层又可以包含多层。每层由多个节点构成, 每层又可以传递给下一层, 直到输出层。除去输入节点, 每个节点都是一个带有非线性激活函数的神经元 (或称处理单元)。隐藏层的神经元与输入层是全连接的, 假设输入层用向量  $X$  表示, 则隐藏层的输出就是  $f(W1X+b1)$ ,  $W1$  是权重 (也叫连接系数),

b1 是偏置, 激活函数  $f$  可以是常用的 sigmoid 函数或者 Tanh 函数。激活函数能够给神经元引入非线性因素, 这使得神经网络可以任意逼近任何非线性函数中, 神经网络可以利用到更多的非线性模型中, 由于 Tanh 函数的均值为 0, 弥补了 sigmoid 函数均值为 0.5 的缺陷, 因此本文使用的是 Tanh 函数作为激活层的函数, 即双曲正切函数, 此函数在原点附近几乎是线性的, 便于下一层的计算, 使得输出结果更加准确快速。

四、实证研究

(一) 数据来源与获取

考虑到房地产评估的特征, 研究的数据主要分为网站上的挂牌交易数据、样本小区信息以及样本地理数据。基于数据获取难度和信息详略程度, 数据来源主要选取有: 链家网站、百度地图 GIS 以及各大房地产中介网站问答板块。

1. 挂牌交易数据: 利用 Python 爬取了链家网站上石家庄市长安区在售二手住宅的相关信息, 共取

得 2 067 条挂牌二手住宅的信息, 通过对部分样本缺失值去除, 筛选重复样本, 去除非典型样本后, 得到了包含全部预期特征变量的 1 045 个样本, 其中包括: 总价、单价、户型、楼龄、总楼层数、物业费、小区户数、装修情况、方位等 9 个指标的相关数据;

2. 样本小区信息: 通过搜查各房地产网站的问答板块核实典型样本的小区信息, 确认了绿化率、容积率, 重点学校划片情况的真实数据;

3. 样本地理数据: 利用百度地图 GIS 测量了样本到市中心的距离, 两公里内学校、医院、地铁、公交、商业中心的数量, 获取了共 6 个指标的相关信息。

(二) 变量选择和量化

通过对房地产评估实务经验总结以及样本分析, 将影响价格的因素主要分为建筑特征、邻里特征、区位特征三个方面。其中总价和单价为目标变量, 其余 16 个为特征变量, 预期会对房价产生一定影响。具体指标及对数据进行量化的方法主要如下:

表 1 变量解释与量化

特征类型	变量名称	变量解释	单位	数据处理
	总楼层数	住宅所在的楼层总数	层	原始数据
	装修情况	精装 5 分, 简装 3 分, 毛坯 1 分		综合比较打分
建筑特征	小区户数	住宅所在小区的总户数, 一定程度能对小区级别进行区分, 如高端社区, 老小区等	户	原始数据
	户型	1 室 0 厅计 1, 1 室 1 厅计分 2, 2 室 1 厅计分 3, 以此类推		综合比较打分
	朝向	朝南计 1, 非朝南计 0		虚拟变量量化
	重点学校	有重点中小学赋值 2, 仅有 1 类赋值 1, 无重点学校赋值 0		虚拟变量量化
邻里特征	绿化率	反映住宅小区的环境情况	—	原始数据
	容积率	反映住宅小区的密集程度	—	原始数据
	到市中心距离	反映住宅城市区位	公里	原始数据
	学校数量	反映周边学校配套情况	个	原始数据
	医院数量	反映周边医院的配套情况	个	原始数据
区位特征	商业中心数量	反映周边商业中心的配套情况	个	原始数据
	地铁数量	反映周边交通情况	个	原始数据
	公交数量	反映周边的交通情况	个	原始数据

对特征变量的量化结束后，对变量进行描述性统计分析，相关统计如表 2：

表 2 描述性统计

变量	样本数	最小值	最大值	平均值	标准差
总价	1 045	52.00	769.0	164.529	89.1468
单价	1 045	5 355.0	30 574.0	15 575.678	3988.261
户型	1 045	1.0	9.0	3.975	1.2261
总楼层	1 045	3.0	38.0	22.615	12.062
方向	1 045	0.0	1.0	0.459	0.4986
装修等级	1 045	1.0	5.0	3.769	1.2151
面积	1 045	32.14	260.17	105.868	36.2143
楼龄	1 045	1.0	38.0	13.000	8.1563
离市中心距离	1 045	1.8	8.5	5.132	1.7497
重点学校	1 045	0.0	2.0	1.025	0.8649
容积率	1 045	0.90	4.47	2.6236	0.8515
绿化率	1 045	0.02	0.70	0.3388	0.0933
物业费	1 045	0.10	2.00	1.1977	0.5959
小区户数	1 045	24.0	7 291	2 543.179	1847.504
地铁数量	1 045	0.0	8.0	2.445	2.0757
公交数量	1 045	1.0	20.0	9.091	1.9996
学校数量	1 045	1.0	20.0	10.810	5.5608
医院数量	1 045	1.0	20.0	4.912	2.7550
商圈数量	1 045	0.0	10.0	3.478	2.4726

(三) 多层感知机 (MLP) 模型实证分析

1. 样本分区

利用 SPSS 内置的多层感知机进行模型预测，根据以往相关文献和研究经验总结，通常以 7：3 的比例划分训练集和检验集。因此在 1 045 个样本中，随机选取 725 个样本进行训练，320 个样本作为测试集对训练出的模型进行检验。

2. 模型检验

平均总体相对误差和相对误差可以评价数据的变化程度，两者值越小，说明评估模型描述数据有更好的精确度，拟合效果更好。通过机器学习后，训练集的平均总体相对误差为 0.068，检验集平均总体相对误差为 0.162，平方和误差也较小，说明本次训练机器学习具有较好的效果，精确度较高。

3. 预测值对比

当评估值和真实值重合得越多，两者就越易汇合到  $y=x$  的直线上，说明模型评估效果越好。如图 2，在多层感知机模型的训练下，单价评估值与实际值重合点占总点数的绝大部分，单价的评估值和真实值的散点图汇聚成为了  $y=x$  的直线，且很清晰。说明模型具有较好的预测效果。

4. 变量重要性

变量重要性通过观察 SPSS 中正态化图例得出。由图 3 正态化图例发现，长安区二手房样本中对因变量影响解释性最强的是楼层数，比例接近 100%，其次是对口重点学校的数量比例超过 80%。这部分

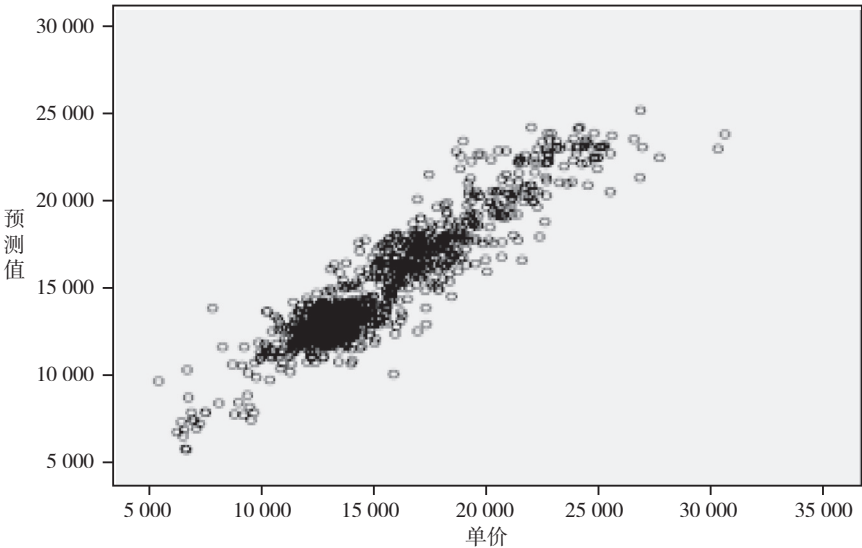


图 2 单价评估值与实际值对比图



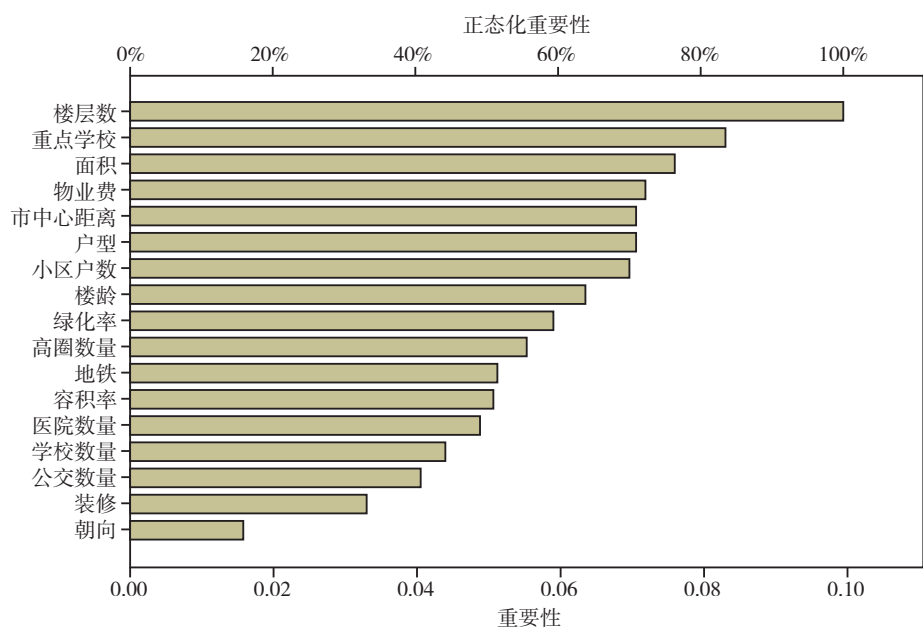


图3 正态化图例

结果符合预期，因为长安区是石家庄市历史较长的行政区域，二手房多为楼龄长的低楼层房屋，在该区后期修建的中高楼层房屋更易受到购买者青睐，此外学区房对口情况也对房价有较大影响力；最弱解释性变量的是朝向，比例仅有10%左右，这一点符合预期，因为网站上挂牌房产绝大部分选择了朝南方位，可能现实中并非如此，大批量的选择朝南使得该指标对房价的影响并不突出；超过60%解释性的因素分别是面积、物业费、市中心距离、户型、小区户数、楼龄。结合面积和物业费的重要性来看，石家庄长安区的二手住宅的房价与住宅的舒适程度，小区

的高档程度有较大关系，在同样的条件下，小区越高端，住宅面积越大，二手住宅的交易价趋于更高。

（四）模型预测效果对比

首先建立参照模型——特征价格模型，通过SPSS计算样本数据后，获得以下特征价格模型系数，其中检验结果中模型调整的 $R^2$ 为0.739，德宾沃森值为2.082，具有较好的拟合效果：如表3。

先随机选取5个样本进行预测值展示，接着对1045个样本进行匹配度对比和误差分析，匹配度反映了真实值和评估值之间的匹配情况，采用评估值除以真实值的方法，计算结果如表4：

表3 特征价格模型系数表

自变量 系数	常数项	户型	总楼层	装修等级	面积	楼龄
	6 939.543	54.678	-85.999	201.265	4.119	-27.107
	容积率	绿化率	物业费	小区户数	地铁数量	公交数量
	652.875	1 751.65	1 516.85	0.174	339.857	-87.881
	离市中心距离	重点学校	商业	学校数量	医院数量	朝向
	-246.035	2 003.730	-144.01	65.572	167.295	287.152

表 4 长安区二手住宅评估值匹配度对比

真实总价	真实单价	评估总价	评估单价	匹配度 1	匹配度 2
特殊价格模型					
101	14 067	78.22	11 828.43	0.7745	0.8408
189.9	14 357	142.0	10 251.60	0.7477	0.7140
202	17 350	146.66	12 677.42	0.7260	0.7306
90	14 129	54.15	11 292.88	0.6017	0.7992
238	18 379	212.94	14 790.38	0.8947	0.8047
多层感知机模型					
101	14 067	94.8	13 428.1	0.9386	0.9545
189.9	14 357	172.4	15 039.8	0.9078	1.0475
202	17 350	187.3	18 272.4	0.9272	1.0531
90	14 129	89.3	13 807.3	0.9922	0.9772
238	18 379	219.4	19 609.7	0.9218	1.0669

在综合 1 045 个样本的匹配度后，多层感知机总价平均匹配度为 1.038，单价的平均匹配度为 1.012，特殊价格模型分别为 0.739、0.724，此外，加入两种模型的误差分析，其中平均绝对误差为模型预测值与真实值的差额，平均相对误差 = ( 预测值 - 真实值 ) / 真实值。由表 5 可知，两种模型在模型评估的精确度方面相差较大，相比之下，多层感知机的评估能力更加突出。

表 5 长安区二手住宅评估效果对比

模型	平均匹配度	平均相对误差	平均绝对误差 ( 元 / 平米 )
特殊价格模型	0.732	-2.14%	-362.2
多层感知机	1.025	1.108%	187.43

为进一步验证多层感知机模型的有效性，根据上述步骤，从石家庄市裕华区 1 500 条挂牌的二手住宅的信息中筛选出包含全部预期特征变量的 570 个样本，分别运用多层感知机模型、特殊价格模型进

行评估，最终得出表 6，由此验证：相比特征价格模型，多层感知机模型评估二手住宅的价值更为准确。

表 6 裕华区二手住宅评估效果对比

模型	平均匹配度	平均相对误差	平均绝对误差 ( 元 / 平米 )
特殊价格模型	1.04578	4.578%	747.93
多层感知机	1.00443	0.443%	72.38

五、结论

通过对房地产批量评估相关文献进行回顾，在前人研究的基础上，较详细地分析神经网络如何在房地产评估中发挥作用，利用爬取的大量样本数据，对石家庄长安区实际挂牌交易的二手住宅进行评估，从理论和实证的角度证明了特征价格模型和神经网络在房地产批量评估有较大使用空间。综合以上，得出的主要结论如下：

第一，变量重要性方面。在多层感知机的学习

下,对石家庄市长安区二手住宅样本而言,可以发现不同变量具有不同的重要性。其中,楼层在所有变量中最重要,占到100%的正态化比例,楼层对房价产生绝对影响,同时其余变量也具有一定的重要性,主要是学区房与否、配套设施等方面也会对房价产生较大影响;

第二,模型训练方面。通过对大量数据的处理和学习后发现,多层感知机在模型训练上具有突出的优势。一方面,训练速度快,即使是上千条数据也能在极短时间内训练得出结果,这为今后建立批量评估的估价系统起到较好的借鉴作用,更体现了批量评估技术在房地产税税基评估中应用优势;另一方面,训练效果好,体现各误差指标值较小,评估值与真实值拟合接近,展出技术预测的优越性。

第三,模型比较方面。经典特征价格模型具有较好的拟合效果,但是与神经网络类的机器学习对比预测效果不太理想。特征价格模型主要是线性形式,函数是提前预设好的,但是实际交易市场中,特征变量与房价并不一定呈线性关系,这就造成误差相对较大。其次,特征变量之间具有一定的相关性,会对最终的结果造成误差。多层感知机则是利用机器模拟神经运行,利用非线性的形式,对样本进行训练,最终得到较好的评估效果。

此外,多层感知机具有较大的算法优化空间,后期还可继续添加算法使其更加优化,比如广义多层感知机<sup>[14]</sup>、BP神经网络的粒子算法、蚂蚁算法、遗传算法、卷积神经网络等等。样本数据有扩充空间,变量可能考虑不全,今后可以完善相关研究。

## 【参考文献】

- [1] “十四五”时期的房产税改革[J].中国注册会计师,2021(03):184-185.
- [2] International Association of Assessing Officer. Standard on Mass Appraisal of Real Property, 2008.
- [3] O'Connor PM. Comparison of Three Residential

Regression models: Additive, Multiplicative, and Nonlinear[J]. Assessment Journal, 2002

- [4] John D. Benjamin, Randall S. Guttery and C. F. Sirmans. Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation[J]. Journal of Real Estate Practice and Education, 2004, 32: 65-77.

[5] 纪益成, 傅传锐. 批量评估: 从价税的税基评估方法[J]. 中国资产评估, 2005.

[6] 孙礼胜, 邓宇, 姚晓艳等. 基于特征价格模型的商品住宅价格影响因素研究——以银川市为例[J]. 宁夏大学学报(自然科学版), 2018, 39(01): 86-92.

[7] 金杰, 罗婷婷. 空间计量模型在房地产批量评估中的应用研究——以昆明市盘龙区二手房挂牌价为例[J]. 中国资产评估, 2021(02): 63-69.

[8] Antipov, EA pokryshevskaya, EB. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics[J]. Expert Systems With Applications, 2012, 39(2): 1772-1778.

[9] 司玺同. 基于机器学习模型的郑州二手房价格评估[D]. 华中师范大学, 2019.

[10] 赵愈, 白晓倩, 许路. 基于BP神经网络的收益性房产税税基批量评估研究[J]. 沈阳建筑大学学报(社会科学版), 2021, 23(02): 151-157.

[11] 陈诗沁, 王洪伟. 基于机器学习的房地产批量评估模型[J]. 统计与决策, 2020, 36(09): 181-185.

[12] 张驰, 郭媛, 黎明. 人工神经网络模型发展及应用综述[J/OL]. 计算机工程与应用: 1-15 [2021-05-19].

[13] 宫釐良, 孙晓峰, 张彦斐. 多层感知器自监督在线修正的道路识别算法[J]. 交通运输系统工程与信息, 2019, 19(4): 101-107.

[14] GARCIA-PEDRAJAS, N, ORTIZ-BOYER. D, HERVAS-MARTINEZ. C. Cooperative coevolution of generalized multi-layer perceptrons[J]. Neurocomputing, 2004, 56: 257-283.

# Application of Multi-layer Perceptron Model in Real Estate Mass Appraisal

Qiu Xiaojie; Luo Rongyun

( School of Finance and Taxation, Hebei University of Economics and Business, Shijiazhuang 050061 )

**Abstract:** The 31st meeting of the Standing Committee of the 13th National People's Congress adopted the decision (Draft) of authorizing the State Council to carry out the pilot work of real estate tax reform in some regions on October 23, 2021. This means that the real estate tax reform has entered a substantive stage. In this context, the real estate value evaluation, which based on mass appraisal, provides technical support for the collection of real estate tax. This paper uses the multi-layer perceptron (MLP) model of artificial neural network in making an empirical study, which takes the second-hand housing in Chang'an District of Shijiazhuang as a sample, and compares the result with the result of Hedonic price model and makes the same empirical analysis on the second-hand housing in Yu hua District. Finally, compared with the classical Hedonic price model, the error between the evaluation results obtained by using the multi-layer perceptron model and the real transaction value is smaller and the matching degree is better.

**Keyword:** Batch Evaluation, Artificial Neural Network, Multilayer Perceptron Model, Hedonic Price Model

