基于粗糙集 BP 神经网络的房产税基批量评估研究

——以广州市为例

李振楠[1] 余炳文[2] 李俊威[2]

（1. 铜陵学院，安徽铜陵 244061；2. 江西财经大学，江西南昌 330013）

【摘　要】自 2021 年发布《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》以来，提出推进房产税立法，健全地方税体系，从法律层面保证房产税实施。房产税基批量评估作为一项繁难的系统工程，基础数据复杂多样，又需兼顾公平性与效率性。借助大数据和计算机智能能够很好地解决此问题，且评估结果科学准确。本文基于粗糙集理论与 BP 神经网络构建房产税基批量评估模型，识别并剔除 7 个对房产税基影响不大的冗余属性，筛选出 20 个显著特征变量，并通过广州市 9 000 套二手普通居民住宅交易样本数据进行验证，结果表明，RS-BPANN 模型在房产税基批量评估中可以实现低成本、高效率、评估结果科学准确的要求，为研究我国房产税基批量评估提供了一定的思路。

【关键词】房产税基　粗糙集　BP 神经网络　批量评估

【中图分类号】F812　　【文献标识码】A　　【文章编号】1007-0265（2023）05-0032-12

一、引言

2021 年发布的《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》中，提出推进房产税立法，健全地方税体系，从法律层面保证房产税的实施。十三届全国人大常委会第三十一次会议决定："授权国务院在部分地区开展房产税改革试点工作。"随着房产税立法的推进，国家不断从法律层面确保房产税实施，我国开征房产税势在必行。大量的房产在征税之前都需要进行价值评估以确定其计税依据。但房产税基影响因素众多且关系复杂，并不是简单的线性关系即可处理，其价值也在不断波动并难以预测。在传统的房产税基批量评估当中，常采用线性模型来构建批量评估模型，评估结果精度不高，委托方较也难接受。其次，发达国家具备相对更加成熟的房产市场，构建的房产税基批量评估方法与评估体系也更加完善。但由于我国房产市场起步较晚，规模更大，交易更频繁，交易种类复杂等特点，照搬国外房产税基批量评估方法与评估体系定不可行，因此研究构建一

［基金项目］江西省教育厅科学技术研究项目《数据资产价值评估研究：产权定价、价值驱动与测度模型》（022120401）阶段性研究成果；江西省教育厅教改研究项目《资产评估专业学位课程框架式案例教学方法优化研究》（JXYJG-2021-098）阶段性研究成果。

［作者简介］李振楠，男，铜陵学院会计学院助教，研究方向：资产评估；余炳文，男，江西财经大学经济学院教授，硕士生导师，研究方向：企业价值评估；李俊威，男，江西财经大学硕士研究生，研究方向：资产评估。

套科学合理，低成本、高效率，具有我国本土国情的房产税基批量评估方法十分必要。

二、基础理论

（一）粗糙集理论

粗糙集理论是一种处理不精确、不确定和不完全数据的数学方法，能在保持分类能力不变的前提下，通过知识约简，导出问题的决策或分类规则。粗糙集理论最核心的是提出属性约简，即可以在保留最基本的信息和保持分类能力不变的前提下，去除重复和冗余的属性或属性值，进而实现对信息的进一步提炼和精简。房产作为商品具有明显的异质性，其属性不同导致价值存在较大差异。在房产税基批量评估当中，利用粗糙集理论深度分析属性的依赖度和重要度，充分识别房产税基影响不明显的冗余属性，挖掘对房产税基有关键性影响的核心属性。以约简后的数据集作为 BP 神经网络的设计依据及训练依据，得到的训练数据表示清晰，训练出来的神经网络不容易出现“过拟合”现象，提高训练效率。其次，伴随着计算机智能和大数据技术的迅速发展，以其为基础的自动批量评估技术日益精进，因粗糙集理论不依赖任何原始数据之外的先验知识或附加条件，在很大程度上与自动批量评估技术相匹配，可以实现低成本、高效率、评估结果科学准确的评估要求。

（二）BP 神经网络

BP 神经网络是一种按照误差逆向传播算法训练的多层前馈神经网络，已经广泛运用于各行各业当中。在整个人工神经网络的发展进程当中，最初的神经网络为单层感知网络，具有计算量小、模型清晰、结构简单等优点。但是随着专家学者研究不断深入，发现它只能解决线性问题，无法处理非线性问题。而后，具有任意复杂的模式分类能力和优良的多维函数映射能力的 BP 神经网络被开发，并在数学上对误差反向传播算法进行完整推导。人工神经网络不需要事先确定含有映射关系的数学方程，而是仅仅通过自身训练或者某种规则，不断接近期望值，因此，人工神经网络实现其功能的核心是算法。BP 神经网络核心算法称为 BP 算法，其核心思想是梯度下降法，即使用梯度搜索技术将该神经网络的实际输出值和期望输出值的误差均方差调整为最小值。BP 算法流程图如图 1 所示，整个计算流程由正向传播和反向传播组成。正向传播从给定输入信号和期望输入开始，经过隐含层逐层处理到达输出层，如若输出层实际输出与期望输出不满足预期要求，则进入反向传播，将误差信号沿原来的连接通路返回，经过对各神经元权值的修改，最终使得误差信号达到最小，符合预期。

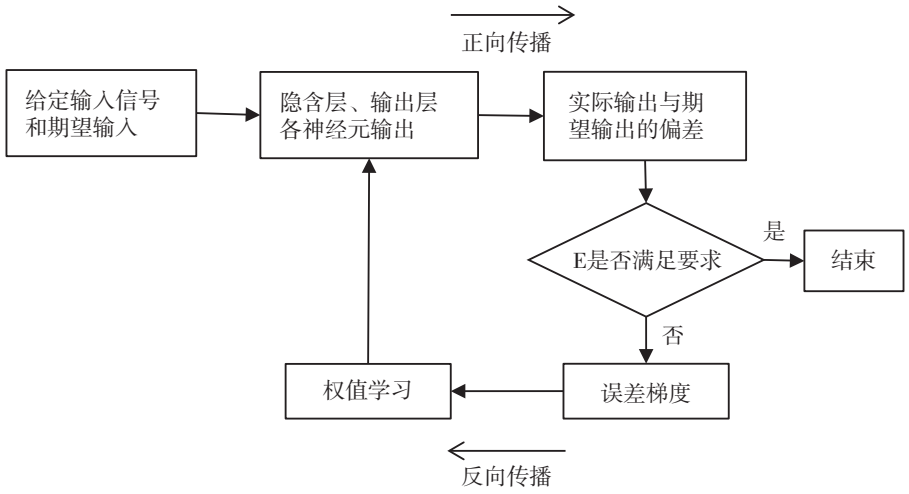

图 1　BP 神经网络的学习传播过程

BP 神经网络现如今已经发展的较为成熟，其最大优点便是具有很强的非线性映射能力和柔性的网络结构。本文首先利用粗糙集理论对房产税基的相关影响因素进行预处理，充分识别对房产税基影响不明显的冗余属性，挖掘对房产税基有关键性影响的核心属性。然后利用简约后的数据集作为 BP 神经

网络的设计依据及训练依据，科学合理确定神经网络的层数以及节点数，结合传递函数，对网络进行充分的训练和建模，直至该神经网络的实际输出值和期望输出值的误差均方差调整为最小。

三、模型构建

（一）评估指标体系的构建

房产税基影响因素众多且关系复杂，其影响主要来源于建筑因素、区位因素和邻里因素三类特征因素。

1. 建筑因素

建筑因素是房产最基础的属性。在考虑建筑因素时，初步选取建筑面积、户型、房屋朝向、装修情况、户型结构、总楼层与所在楼层、电梯与梯户比例、建筑类型、建筑结构、建成年代作为特征变量。

①建筑面积：主要是指套内建筑面积，即套内房屋使用空间的面积，其大小决定了房产的可使用空间大小，是房产最重要属性之一。

②户型：主要指房屋结构形态，一套房屋通常包括客厅、厨房、阳台、卧室、餐厅、卫生间等空间，空间的数量、面积大小和位置都会对房产价值造成影响。考虑到相关数据的采集难易程度，本文将选取卧室、客厅、厨房、卫生间的数量作为特征变量。

③房屋朝向：主要指房屋采光面最大的方向。一般来说，朝南方向房产价值比同等户型其他朝向价值高。

④装修情况：主要指在一定区域和范围内进行，包括墙体、天花板、地板、水电施工、景观等所实现，依据美观规则和一定设计理念形成的一整套施工和设计方案。根据装修档次的高低，可以分为精装、简装、毛坯和其他。

⑤户型结构：主要指各个房间的空间布局情况。根据房屋内厅、卧、卫、厨、阳台处于几个高度不同平面的情况，可以将房屋分为平层、错层、复式和跃层等。

⑥总楼层与所在楼层：总楼层是指一栋建筑物的总楼层数，所在楼层是指所处的楼层数，可以简单分成低楼层、中楼层和高楼层三个档次。

⑦电梯与梯户比例：主要指建筑物是否配备电梯、电梯数和每层楼住户数比例。梯户比越高，说明一栋楼人口居住密度越大，等电梯的时间也随之延长。因此配备有电梯并且梯户比例越低的住房，其房产价值将会更高。

⑧建筑类型：主要分为平房、塔楼、板楼和塔板结合。

⑨建筑结构：主要指在房屋建筑中，由屋架、板、梁、柱等各种构件组成的能够承受各种作用的体系。根据所用材料，可以分为钢混结构、钢结构、砖混结构、砖木结构以及混合结构。

⑩房屋年龄：主要指自房屋竣工验收合格交付使用之日起开始计算至评估基准日的年份数。

2. 区位因素

区位因素是影响房产价值的关键性因素，它反映了周边事物与被评估房屋的空间位置关系。

①所处行政区：对于具有极强区域性的房产来说，其价值会因地域表现出巨大差异，进而影响房产税基的价值。在一定区域间，政治、经济、文化等发展水平基本保持一致，所构建的房产税基批量评估模型也更加合理准确。

②交通便利程度：主要是指房产周围交通设施的发达程度，交通条件越好的地区，区域内的房产价值也会相对较高。

③至城市中央商务区的距离：中央商务区一般处于一个城市的腹地中心，周边拥有良好的交通、购物、医院、学校、金融等资源，因此房产价值也相对较高。

3. 邻里环境因素

邻里环境因素主要是指房产周围的、对家庭生活产生影响的特征，主要包括医疗、教育、商场、公园等基础设施方面和小区绿化率、容积率、管理费、车位数量等方面。

①基础配套设施资源：主要指包括通讯、城市供水供电、提供无形产品以及科教文卫等外部设施。基础配套设施越完善，该住宅小区的房屋就越受购房者的喜爱，房产价值也就越高。

②教育设施配套资源：主要是指对教育资源的关注，包括中小学以及幼儿园教育等。目前我国很多地方均将教育资源与房产价值挂钩，房产周围的教育设施配套资源已成为影响其价值的重要因素。

③住宅小区环境：一个管理有序、绿化优美的

小区内的房产价值相对来说会更高。因此，本文除了选取小区绿化率、容积率和房屋总数量等变量来直观衡量小区环境的优劣之外，还将选取小区物业费这一变量来侧面反映小区的档次和管理水平。

4. 变量的分类及量化方法

为确保样本数据质量，在选择房产税基影响因素时，主要从四个方面进行考虑：影响因素对房产税基价值的影响是否显著、影响因素之间是否存在共线性、应税房产样本数据是否可信以及是否能够客观量化。经过综合考虑，初步选取的特征变量为建筑面积、室数量、厅数量、厨数量、卫数量、房屋朝向、装修情况、房屋年龄、户型结构、总楼层、所处楼层、建筑类型、梯户比例、建筑结构、所处行政区、电梯设备、公交站数、地铁站数、城市中央商务区距离、生活医疗设施、基础配套、环境景观、配套教育、绿化率、容积率、房屋总数、物业管理费。所考虑的变量分类说明以及量化方法如表 1 所示。

表 1　变量的分类说明及量化方法

特征分类	变量	变量名称	量化方法
建筑因素	X_1	建筑面积	以平方米为单位的住宅实际面积数
	X_2	卧室	住宅内卧室的数量
	X_3	客厅	住宅内客厅的数量
	X_4	厨房	住宅内厨房的数量
	X_5	卫生间	住宅内卫生间的数量
	X_6	房屋朝向	朝南为 5；朝东南、西南为 4；朝东、西为 3；朝东北、西北为 2；朝北为 1
	X_7	装修情况	精致装修为 4；普通装修为 3；毛坯为 2；其他为 1
	X_8	户型结构	跃层为 4；错层为 3；复式为 2；平层为 1
	X_9	总楼层	住宅所在幢的总层数
	X_{10}	所在楼层	住宅层数 2/3 以上的楼层为 5；1/3–2/3 的楼层为 3；1/3 以内的楼层为 1
	X_{11}	电梯	配备有电梯为 1；没有配备电梯为 0
	X_{12}	梯户比例	每层楼三户及以下为 5；四到六户为 4；七到九户为 3；十到十二户为 2；十二户以上为 1
	X_{13}	建筑类型	塔板结合为 4；板楼为 3；塔楼为 2；平房为 1
	X_{14}	建筑结构	混合结构为 4；钢混结构为 3；框架结构为 2；砖混结构为 1
	X_{15}	房屋年龄	根据住宅的建成年代计算，2020 年建成的房屋年龄为 1；2019 年为 2；以此类推
区位因素	X_{16}	所处行政区	住宅所处的行政区
	X_{17}	公交站数量	住宅周围 1km 内的公交站数量
	X_{18}	地铁站数量	住宅周围 1km 内的地铁站数量
	X_{19}	至城市中央商务区距离	住宅至城市中央商务区的直线距离（千米）
邻里环境因素	X_{20}	医院数量	住宅周围 1km 内的医院数量
	X_{21}	商场数量	住宅周围 1km 内的商场数量
	X_{22}	公园数量	住宅周围 1km 内的公园数量
	X_{23}	学校数量	住宅周围 1km 内的重点学校数量
	X_{24}	绿化率	绿化用地面积与总用地面积之比
	X_{25}	容积率	总建筑面积与净用地面积的比率
	X_{26}	物业费	住宅所在小区的物业费（元 / 平方米 / 月）
	X_{27}	房屋总数	住宅所在小区的房屋总数量

（二）模型构建

1. 输入输出的确定

BP 神经网络的输入的选择为经过粗糙集理论筛选过后的特征变量，输出的选择为待评估房产每平方米的成交价格。

2. 属性约简

将粗糙集属性约简知识运用于房产税税基评估指标体系构建中，就是将房产价值评估指标体系中的指标作为条件属性集，将房产市场价值作为决策属性，约简去对房产市场价值没有影响或者影响不大的指标，求解相对约简属性集，约简后的属性集即为简化后的房产市场价值评估指标体系。具体步骤如下：

①建立决策表 $D=\{U,R,V,f\}$ 和条件属性集 $C=\{\alpha_1,\alpha_2,\cdots,\alpha_i\}$；

②计算出 C 相对于 D 的核心属性 $core_D(C)$；

③令 $B=core_D(C)$，对于任意非核属性 $\alpha_i^*\in D-B$，属性重要度计算如下：

$$sig(\alpha_i^*,B;D)=\frac{card(pos_{B+|\alpha_i^*|}(D))-card(pos_B(D))}{card(pos_B(D))} \quad (1)$$

④比较属性重要度，删除对房产市场价值没有影响或者影响不大的属性，得出约简属性集。

3. BP 神经网络构建

（1）确定网络层数

神经网络可以包含单个隐含层，也可以包含多个隐含层。单个隐含层可以通过增加神经元节点个数的方式来实现所需非线性映射，只有函数间断时才需要利用两个隐层达到网络训练目的，具有单隐层的前馈网络完全可以满足所有连续函数映射关系。因此，本文采用单个隐含层来构建房产税批量评估模型。

（2）确定各层节点数

确定 BP 神经网络中的各层节点数必须从问题入手。对于输入层和输出层，输入层只接受外部数据输入，故外部输入向量维数决定了输入层节点数，该节点数与经粗糙集理论约简后变量个数相等。输出层仅仅输出房产税基评估结果，因此输出层节点数确定为一个。确定隐含层节点数复杂且十分重要，对构建的神经网络模型性能有重大影响，可能直接导致训练时出现“过拟合”的情况。当确定隐含层节点数太少时，可能直接导致 BP 神经网络训练失败或者达不到预期效果。当确定隐含层节点数太多时，可能导致 BP 神经网络训练陷入局部极小点，进而难以得到最优值。通过下列公式使用试凑法不断调节 α 值逐渐增加隐含层节点数可以确定隐含层最佳节点数。

$$x=[\sqrt{m+n}]+\alpha \quad (2)$$

其中：x 为隐含层节点数，m 为输入节点数，n 为输出节点数，[x] 为取整函数，α 为常数。

（3）节点间链接

本文只考虑上下神经元之间的连接，同一层级之间的节点不存在连接。

（4）传递函数

本文选取 Sigmoid 函数作为神经网络的激活函数。Sigmoid 函数作为生物学中一个常见的函数，表现为 S 型，可以分为 Log-Sigmoid 函数和 Tan-Sigmoid 函数。

Log-Sigmoid 函数表现为：

$$S(x)=\frac{1}{1+e^{-x}} \quad (3)$$

Log-Sigmoid 函数的导数可以用其自身表示：

$$S'(x)=S(x)(1-S(x)) \quad (4)$$

Tan-Sigmoid 函数表现为：

$$S(x)=\frac{2}{1+e^{-2x}}-1 \quad (5)$$

Tan-Sigmoid 函数的导数可以用其自身表示：

$$S'(x)=1-S(x)\times S(x) \quad (6)$$

这两个函数都是连续、单调递增的数值函数，常被应用于基于 BP 算法的神经网络中。Log-Sigmoid 函数的值域为（0,1），Tan-Sigmoid 函数的值域为（-1,1）。一般情况下，BP 神经网络隐含层的传递函数是 S 形函数，输出层是线性函数。当然，输出层也可采用 S 型函数，若输出层为 S 型函数，则输出值的范围为该 S 型函数的值域。利用 S 形函数或其导数可以求得 BP 神经网络里某个神经元的总和、目标值和误差值等。

（5）学习算法

BP 神经网络核心算法称为 BP 算法，其核心思想是梯度下降法，即使用梯度搜索技术将该神经网络的实际输出值和期望输出值的误差均方差调整为

最小值。本文采用的 BP 神经网络由输入层、隐含层、输出层所组成；外部输入向量维数决定输入层节点数，输出层节点数确定为一个，隐含层节点数为若干个。首先，通过输入层输入经粗糙集理论约简后的训练样本，设置因变量、相关自变量、训练最大误差以及最大训练次数等相关参数。其次，输入经粗糙集理论约简后的训练样本经过隐含层最后到达输出层。最后，将输出层的输出值与实际样本输出值进行比较，交替重复上述过程，使得误差达到预计最小值，即完成整个学习过程。

在 BP 神经网络中，有正向传播和反向传播两种方式。正向传播从给定输入信号和期望输入开始，经过隐含层逐层处理到达输出层，如若输出层实际输出与期望输出不满足预期要求，则进入反向传播，将误差信号沿原来的连接通路返回，经过对各神经元权值的修改，最终使得误差信号达到最小，符合预期。图 2 描绘了一层神经元产生的一组函数信号$y_i(n)$反馈给神经元 j，m 为作用于神经元 j 不包括偏置的所有输入个数，突触权值$w_{j0}(n)$等于神经元 j 的偏置b_j。神经元 j 的激活函数φ输入处产生的诱导局部域为$v_j(n)$，输出处的函数信号为$y_j(n)$，期望输出为$d_j(n)$，误差信号为$e_j(n)$。

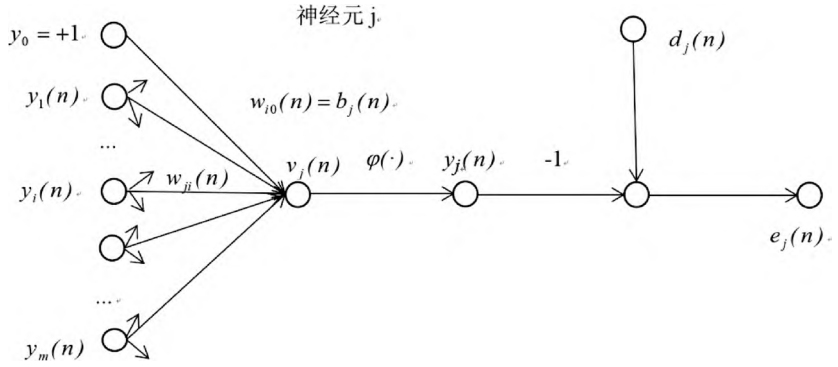

图 2　神经元 j 传播

①前向传播

在神经元 j 的激活函数输入处产生的诱导局部域$v_j(n)$，即：

$$v_j(n)=\sum_{i=0}^{m}w_{ji}(n)y_i(n) \tag{7}$$

假定φ_j是激活函数，则出现在神经元 j 输出处的函数信号$y_j(n)$为：

$$y_j(n)=\varphi_j(v_j(n)) \tag{8}$$

②误差反向传播

在图 2 中，神经元 j 的实际输出为$y_j(n)$，期望输出为$d_j(n)$，则输出所产生的误差信号为：

$$e_j(n)=d_j(n)-y_j(n) \tag{9}$$

上式中，$d_j(n)$为期望响应向量$d(n)$的第 j 个元素。

将均方根差最小化，使函数连续可导，则神经元 j 的瞬时误差能量为：

$$E_j(n)=\frac{1}{2}e_j^2(n) \tag{10}$$

相加所有输出层神经元误差能量，可得整个网络的全部瞬时误差能量为：

$$E_n=\sum_{j\in C}E_j(n)=\frac{1}{2}\sum_{j\in C}e_j^2(n) \tag{11}$$

上式中，集合 C 为输出层的所有神经元。

BP 算法通过反复修正权值使式（11）E_n最小化，采用梯度下降法对突触权值$w_{ji}(n)$应用一个修正值$\Delta w_{ji}(n)$，正比于偏导数$\frac{\delta E(n)}{\delta w_{ji}(n)}$。根据微分链式规则，可以把这个梯度表示为：

$$\frac{\partial E(n)}{\partial w_{ji}(n)}=\frac{\partial E(n)}{\partial e_j(n)}\frac{\partial e_j(n)}{\partial y_j(n)}\frac{\partial y_j(n)}{\partial v_j(n)}\frac{\partial v_j(n)}{\partial w_{ji}(n)} \tag{12}$$

偏导数$\frac{\delta E(n)}{\delta w_{ji}(n)}$代表一个敏感因子，决定突触权值$w_{ji}$在权值空间的搜索方向。

在式 (11) 两边对$e_j(n)$取微分，得到：

$$\frac{\partial E(n)}{\partial e_j(n)}=e_j(n) \tag{13}$$

在式（9）两边对$y_j(n)$取微分，得到：

$$\frac{\partial e_j(n)}{\partial y_j(n)}=-1 \tag{14}$$

在式（8）两边对$v_j(n)$取微分，得到：

$$\frac{\partial y_j(n)}{\partial v_j(n)} = \varphi_j'(v_j(n)) \quad (15)$$

在式（7）两边对$w_j(n)$取微分，得到：

$$\frac{\partial v_j(n)}{\partial w_{ji}(n)} = y_j(n) \quad (16)$$

将式（13）至式（16）代入式（12）可得：

$$\frac{\partial E(n)}{\partial w_{ji}(n)} = -e_j(n)\varphi_j'(v_j(n))y_i(n) \quad (17)$$

应用于$w_{ji}(n)$的修正$\Delta w_{ji}(n)$定义为：

$$\Delta w_{ji}(n) = -\eta \frac{\partial E(n)}{\partial w_{ji}(n)} \quad (18)$$

上式中，η为误差反向传播的学习率，负号表示在权空间中梯度下降。

将式（17）代入式（18）可得：

$$\Delta w_{ji}(n) = \eta \delta_j(n) y_j(n) \quad (19)$$

其中$\delta_j(n)$为局部梯度：

$$\delta_j(n) = -\frac{\partial E(n)}{\partial e_j(n)} \frac{\partial e_j(n)}{\partial y_j(n)} \frac{\partial y_j(n)}{\partial v_j(n)} = e_j(n)\varphi_j'(v_j(n)) \quad (20)$$

局部梯度指明了突触权值所需要的变化。

（6）学习率的确定

学习率表示每次参数更新幅度大小，会影响系统学习过程的稳定性。学习率设置过大，提升网络训练前期学习效率，使整个模型更快接近最优解，但可能跨过或忽略最小值，导致一直来回震荡而无法收敛。学习率设置过小，影响隐含层到输出层的连接权值、输入层到隐含层的连接权值，导致迭代更新速度缓慢。因此，本文为保证学习过程收敛性倾向选取较小的学习率，在 0.01 到 0.8 之间。

（7）初始权值的确定

初始权值的选取与神经网络预测能力有着密切联系。选择不同初始权值，得出的神经网络性能也不一样。BP 算法对于初始权值选取较为敏感，一般要求网络初始值分布在 -1 到 1 之间。

（8）期望误差的确定

在设计网络的训练过程中，网络误差经过样本多次迭代计算后将达到期望误差，并对多个不同期望误差值的网络进行训练，而后进一步通过综合因素来确定合适的期望误差。

四、实证分析

（一）数据来源

在“房住不炒”的大环境下，所有政策调控都在不断升级，取得的效果也是很明显的。2021 年，广州市印发《广州市人民政府办公厅关于进一步促进房地产市场平稳健康发展的意见》，将重拳整治房产市场秩序，着力稳地价、稳房价、稳预期，促进广州房产市场平稳健康发展。根据《中国房产服务行业消费者满意度调查报告》显示，贝壳网与链家网关于房产中介真房源可信度和真房源满意度上在同行业中均位列前茅。鉴于 BP 神经网络分析需要大量数据，从贝壳网和链家网上搜集广州市二手居民住宅小区的房产成交数据，获取样本小区房屋成交时的实际成交价和相关信息获取各小区名称、建筑年代、装修、朝向、物业费等特征。其次，利用高德地图等来获取样本房产交通、区位、基础设施配套等数据，主要包括：重点学校数量、地铁站数、公交站数、周围配套设施、至城市中央商务区距离等。

（二）变量的筛选

依据粗糙集理论，搜集 2021 年 1 月至 6 月广州市 9 100 套二手居民住宅成交数据进行预处理。其中，全域 U={ 所有样本 }；决策属性 D={ 成交价格 }；条件属性集 C={ α_1，α_2，…，α_i }={ 建筑面积，卧室，…，房屋总数 }。具体处理步骤如下。

第一步，用 SPSS 软件的 k-means 聚类方法，以欧氏距离计算一些连续变量属性的相似度，将两两相似度大的对象归于同一类，并用将实际连续数值分成 5 份，实现连续属性离散化。再对新的决策表进行属性重要度分析，变量离散化结果如表 2 所示。

表 2　部分变量的聚类中心

房价（元 / 平方米）	低	中低	中	中高	高
	19 866.77	31 648.32	544 308.85	61 604.89	92 231.82
建筑面积（平方米）	小	中小	中	中大	大
	56.05	87.98	129.16	208.22	331.46

续表

总楼层（层）	低	中低	中	中高	高
	2	15	28	41	55
建筑年代（年）	很新	新	一般	旧	很旧
	8	16	23	29	40
至城市中央商务区距离（千米）	近	中近	中	中远	远
	0.13	0.28	0.44	0.65	0.88
绿化率（百分比）	低	中低	中	中高	高
	17.13	31.74	41.30	53.04	74.25
容积率	低	中低	中	中高	高
	2.49	5.43	9.56	16.77	25.91
物业费（元/平方米/月）	低	中低	中	中高	高
	0.81	1.71	2.72	3.91	7.16

第二步，求取 C 相对于 D 的核心属性 $core_D(C)$。由粗糙集理论可知，对于每一个属性α_i，若存在$pos_{ind(C)}(D)=pos_{ind(C-\alpha i)}(D)$，则认为$\alpha_i$在 C 中对 D 是不必要的，反之则认为$\alpha_i$在 C 中对 D 是必要的。用 Rosetta 软件的 Johnson 算法进行相对核心的约简计算，并得到相对核心属性，结果如图 3 所示：

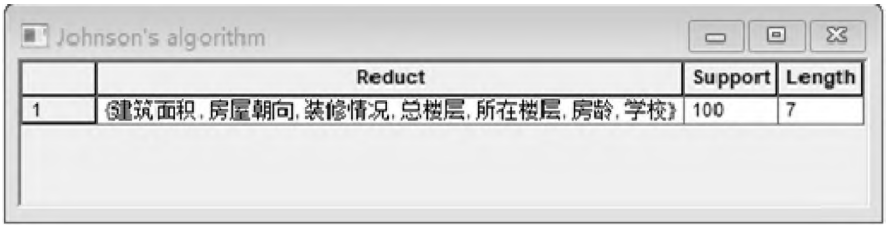

图 3　Johnson 算法输出结果

由图 3 可以得到，决策表的核心属性 $core_D$（C）=｛建筑面积，房屋朝向，装修情况，总楼层，所在楼层，房龄，学校数量｝。

第三步，计算非核属性的重要度。设 B=$core_D$（C），α_i^*是非核属性，经过 Rosetta 软件计算后，得到每个非核属性的重要度如表 3 所示。

表 3　各属性的重要度

条件属性集	属性重要度	是否作为评估指标
建筑面积	core	是
卧室	0.32	是
客厅	0.14	是
厨房	0.03	否
卫生间	0.11	是

续表

条件属性集	属性重要度	是否作为评估指标
房屋朝向	core	是
装修情况	core	是
户型结构	0.02	否
总楼层	core	是
所在楼层	core	是
电梯	0.23	是
梯户比例	0.15	是
建筑类型	0.12	是
建筑结构	0.03	否
房龄	core	是
所处行政区	0.26	是
公交站数量	0.19	是

续表

条件属性集	属性重要度	是否作为评估指标
地铁站数量	0.06	否
至城市中央商务区距离	0.15	是
医院数量	0.19	是
商场数量	0.02	否
公园数量	0.14	是
学校数量	core	是
绿化率	0.18	是
容积率	0.06	否
物业费	0.27	是
房屋总数	0.05	否

第四步，确定约简属性集。从表 3 可以得出厨房、户型结构、建筑结构、地铁站数量、商场数量、容积率、房屋总数这 7 个变量属性重要度相对较小，数值近似等于 0，则该指标在房产税基批量评估中是冗余评估指标，可以直接剔除。其余指标明显大于 0，说明它们在评估中起到重要作用，因此最后确定的条件属性集 = {建筑面积，室，厅，卫，房屋朝向，装修状况，总楼层，所在楼层，电梯，梯户比例，建筑类型，房龄，所处行政区，公交站数量，至城市中央商务区距离，医院数量，公园数量，学校数量，绿化率，物业费}，并且将它们作为神经网络的输入层结点。

（三）模型数据的输入

为满足 BP 神经网络训练需求，将借助 Matlab 软件实现房产税基批量评估模型的构建、训练与检验。本文收集了 9 100 套二手房交易样本数据，选取其中 9 000 个样本数据用以模型的训练，剩下 100 个样本数据则作为检验模型的测试样本。

（四）模型数据的归一化处理

为避免因输入输出的变量数量级不同造成对神经网络预测结果准确性产生影响，故需要对输入输出参数归一化处理，使变量初始值分布在 -0.5 到 0.5 之间。

（五）BP 神经网络模型的建立

1. 输入神经元数

根据粗糙集理论对变量筛选，以广州市二手住宅为研究对象房产税基评估价值的显著指标个数为 20 个，因此输入层的神经元数设定为 20 个。

2. 输出神经元数

输出层变量为房产的市场交易价值，因此只需要 1 个输出层神经元。

3. 隐含层节点数

通过多次拟合测试结果分析，将不同节点数代入模型进行训练得到拟合优度的结果如表 4 所示。当隐含层节点数为 25 个时，模型拟合程度最高。

表 4　不同隐含层节点数测试结果

节点数	1	2	3	4	5	均值
18	0.9081	0.9055	0.9172	0.9058	0.9069	0.9087
19	0.9027	0.9016	0.9068	0.9047	0.8953	0.9022
20	0.9081	0.9117	0.9092	0.9174	0.9136	0.9120
21	0.9150	0.9145	0.9135	0.9109	0.9163	0.9140
22	0.9192	0.8882	0.9150	0.9145	0.9173	0.9108
23	0.9152	0.9166	0.9216	0.9173	0.9168	0.9175
24	0.9191	0.9228	0.9218	0.9209	0.9184	0.9206
25	0.9205	0.9272	0.9209	0.9231	0.9247	0.9233
26	0.9184	0.92184	0.9171	0.9144	0.91881	0.9181
27	0.9247	0.9221	0.9198	0.9176	0.9211	0.9211

4. 传递函数

本文采用 Sigmoid 函数作为传递函数，在其他参数和样本数据不变的前提下对 Tan-Sigmoid 和 Log-Sigmoid 函数分别进行测试，结果得表 5.

表 5　不同传递函数测试结果

传递函数	1	2	3	4	5	均值
tansig	0.9221	0.9271	0.9247	0.9224	0.9223	0.9237
logsig	0.9197	0.9198	0.9160	0.9171	0.9182	0.9182

依据表 5 可得，在房产税基批量评估当中 Tan-Sigmoid 函数的准确率更好，故选择 Tan-Sigmoid 函数作为传递函数。

5. 训练函数

分别对 BP 神经网络权值调整算法中 trainlm 和 trainbr 算法模型进行测试，结果得表 6。

表 6　不同训练函数测试结果

训练函数	1	2	3	4	5	均值
trainlm	0.9063	0.9083	0.9018	0.8783	0.8997	0.8989
trainbr	0.9188	0.9216	0.9186	0.9231	0.9248	0.9214

依据表 6 可得 trainbr 算法模型的准确率更好，故权值调整算法采用 trainbr 算法进行训练。

6. 误差函数

神经网络输出与期望输出的误差可以采用误差平方和性能函数 mse 进行计算。根据以上要求，建立一个未经训练的 BP 神经网络模型。并且通过 Mathlab 软件随机挑选出 20% 的训练样本用于对模型的预检验。

（六）模型的训练

1. 相关参数的确定

（1）训练次数

由于样本数量偏大，为使神经网络得到充分的训练，因此将最大训练次数设定为 200 次。

（2）训练误差

本文选择系统默认的均方误差（mse）作为神经网络训练的目标误差，并将其设定为 0.0001。

（3）学习率

为了防止训练过程出现波动徘徊现象的发生，通过多次训练，最终发现当学习率为 0.01 时，神经网络的训练效率和训练效果相对最好。因此将学习率设定成 0.01。

（4）显示间隔

为方便观察训练过程以及结果，设定每迭代为 10 次，系统便自动记录并反馈神经网络的训练程度。

2. 神经网络的训练

相关系数确定好以后，便开始训练神经网络。通过对 9 000 个样本的迭代，神经网络训练结果如图 4 所示。

图 5 为神经网络误差变化图，性能为程序中指定的训练结束参数，当期望误差满足结束条件时，网络训练就会结束。其次明显可得知，在迭代次数达到 50 次左右时，曲线开始接近于一条直线，模型已经开始收敛。网络中指定期望误差为均方误差，其初始误差为 0.271，结束值为 0.00204，经 200 步计算虽然未达到预先设定期望误差 0.0001，但最优误差也收敛到了 0.00204，并且收敛速度快，也没有出现过度拟合情况。

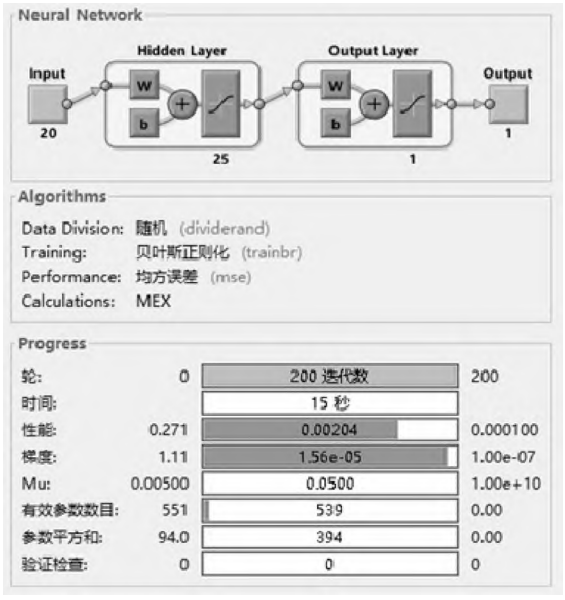

图 4　神经网络训练过程与结果

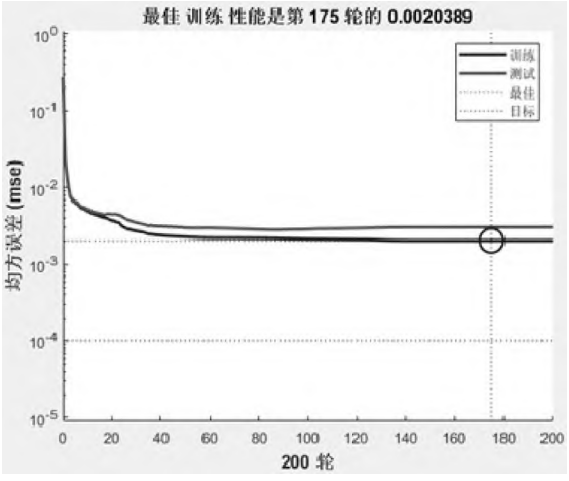

图 5　神经网络误差变化图

如图 6 所示，在大样本数据之下，绝大多数样本数据都集中在拟合函数附近，并且总体拟合度达到 0.92443，说明 BP 神经网络训练得非常成功，训练结果误差小，收敛快。

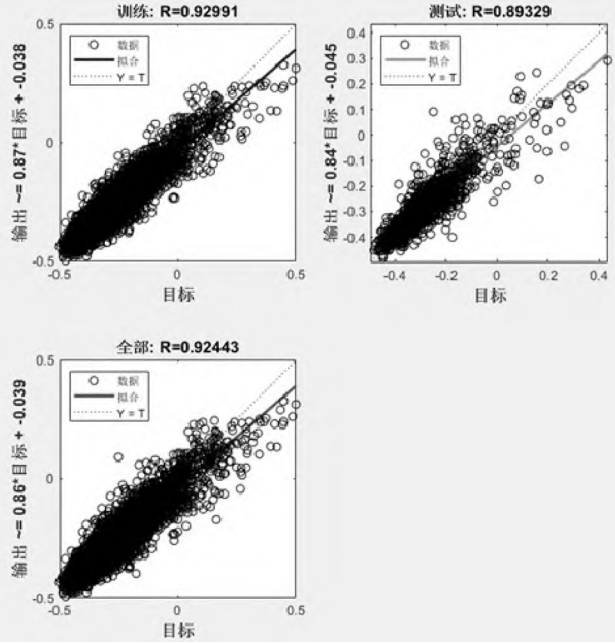

图 6　神经网络回归结果

（七）测试样本检验

最后选取 100 组样本验证训练以后神经网络的准确性，将测试样本的因变量进行归一化处理，导入已经训练完成的神经网络中得出评估值，并与实际价格进行比较，并计算其绝对误差率和相对误差率。测试结果如图 7、图 8、图 9。

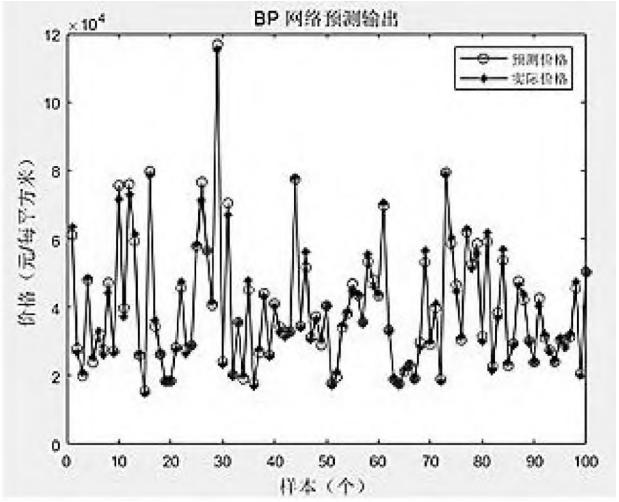

图 7　BP 神经网络预测输出与实际输出对比

从测试结果可以看出，通过训练好的 BP 神经网络所得到的输出值与真实成交价格比较接近，说明该模型符合实际，具有一定的实践意义。

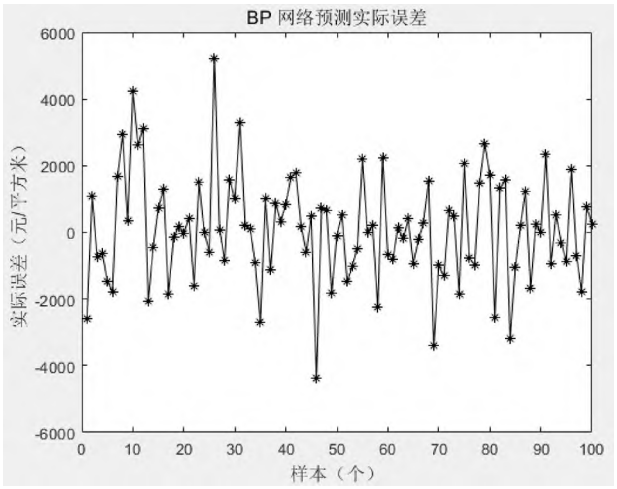

图 8　BP 神经网络实际输出误差

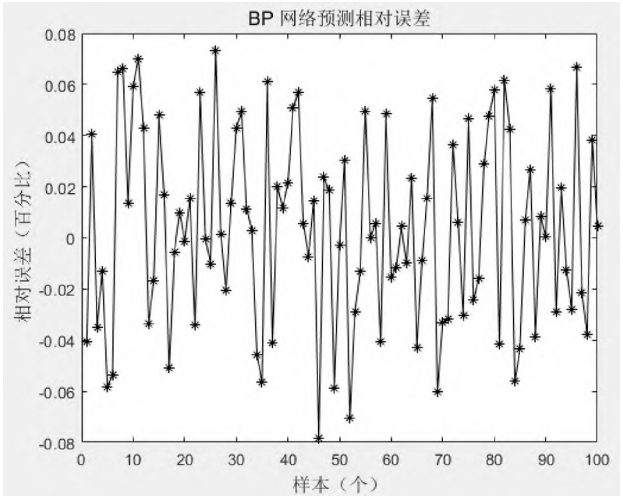

图 9　BP 神经网络相对误差

五、总结

房产税是对所有房产交易和保有环节征收的一个税种。房产税基影响因素众多且关系复杂，并不是简单的线性关系即可处理，并且，现如今理论及实务界均未给出一个具体的函数模型来准确地表达房产税基与其影响因素之间的映射关系。BP 神经网络具有同时处理多个因素和条件的信息问题的自适应和自学习能力。本文通过粗糙集理论深入分析属性依赖度和重要度，识别并剔除 7 个对房产税基影响不大的冗余属性，挖掘出对决策属性有关键性影响的 20 个核心属性。而后将该 20 个核心影响因素量化值作为神经网络的输入，样本的价格作为网络的期望输出，进行网络训练，确定其结构与参数，建立起房产税基与其影响因素的非线性关系模型。

并通过广州市 9000 套二手普通居民住宅交易样本数据进行验证，结果表明，RS-BPANN 模型在房产税基评估方面具有优秀的性能，可以实现低成本、高效率的评估要求，为研究我国房产税基批量评估提供了一定的思路。

【参考文献】

[1] 赵丙奇 , 林圣豪 , 章合杰 . 房产税对不同类型住房价格的影响——来自重庆房产税试点的证据 [J]. 社会科学战线 ,2022(07):88-97.

[2] 张苏 . 房产税改革中税基批量评估研究进展 [J]. 特区经济 ,2021(05):154-156.

[3] 赵愈 , 许路 . 基于 BP 神经网络的房产税税基评估研究 [J]. 沈阳建筑大学学报 (社会科学版),2020,22(02):144-150.

[4] 王阿忠 , 李倩 . 基于粗糙集神经网络的房产税基批量评估研究 [J]. 福州大学学报 (哲学社会科学版),2019,33(05):30-37.

[5] 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.

[6] 贾万龙 , 王万雄 . 基于 BP 神经网络的合肥都市圈 GDP 预测 [J]. 西安电子科技大学学报 (社会科学版),2022,32(01):37-44.

Research on Batch Assessment of Property Tax Base Based on Rough Set Theory and BP Neural Network

——Take Guangzhou City as an Example

Li Zhennan[1], Yu Bingwen[2], Li Junwei[2]

(1.Tongling University; 2.Jiangxi University of Finance and Economics)

Abstract: Since the release of the Outline of the 14th Five-Year Plan and 2035 Vision for National Economic and Social Development of the People's Republic of China in 2021, it has proposed to promote the property tax legislation, improve the local tax system, and ensure the implementation of the property tax from the legal level. As a complicated system engineering, the bulk evaluation of property tax base has complex and diverse basic data, and it needs to take into account both fairness and efficiency. This problem can be well solved with the help of big data and computer intelligence, and the evaluation results are scientific and accurate. Based on rough set theory and BP neural network, this research builds a batch evaluation model of property tax base, identifies and eliminates 7 redundant attributes that have little impact on the property tax base, screens out 20 significant characteristic variables, and verifies them through the sample data of 9 000 second-hand ordinary residents' housing transactions in Guangzhou. The results show that, RS-BPANN model can achieve low cost, high efficiency and scientific and accurate evaluation results in the batch evaluation of property tax base, which provides a certain idea for studying the batch evaluation of property tax base.

Keywords: Property tax base，Rough set theory，BP neural network，Mass appraisal