文章编号:2095-0365(2020)02-0014-09

基于多源数据的房地产价格指数预测模型及实证研究

朴春慧[1], 武旭晨[1,2], 蒋学红[3], 李玉红[4]

(1.石家庄铁道大学 信息科学与技术学院,河北 石家庄 050043;
2.中国银行河北省分行 信息科技部,河北 石家庄 050000;
3.河北省住房和城乡建设厅 信息中心,河北 石家庄 050051;
4.石家庄铁道大学 经济管理学院,河北 石家庄 050043)

摘 要:房地产的价格变化对社会经济发展有显著的影响,准确预测房地产市场价格变化并对其进行有效调控显得尤为重要,但使用房价作为评估房地产市场的度量指标有一定的局限性。住宅销售价格指数是由国家统计局发布的综合反映住宅商品价格水平总体变化趋势和变化幅度的相对数,为探讨新建商品房住宅销售价格指数的预测方法及其预测有效性,利用与相关的房地产供求关系、社会宏观经济指标、国家货币政策和民众对房价的预期等多源数据,构建了一套房地产价格指标体系。分别使用BP-Adaboost和支持向量回归机两种机器学习算法构建房地产评估模型,同时设计了一个调参算法对支持向量回归机模型进行参数优化。在实证中使用华北某市的房地产月度数据对两种模型进行训练和预测,并与ARIMA模型和经典BP神经网络模型做对比。实验结果表明,BP-Adaboost模型的预测误差最小,使用BP-Adaboost模型预测房地产价格指数具有可行性。

关键词:房地产预测;BP-Adaboost算法;支持向量回归机;住宅销售价格指数
中图分类号:F293.35 **文献标识码**:A **DOI**:10.13319/j.cnki.sjztddxxbskb.2020.02.02

一个国家或地区的经济发展与房地产业的运行状况息息相关。但当前与房地产价格相关的理论研究不成熟,评估方法较为依赖对过去的经验,一定程度上影响了房地产评估行业的发展[1]。当前对房地产价格的预测存在着几点不足。第一,由于我国房地产市场起步较晚,数据不完整,现有研究大都以年度数据对模型进行实证研究,数据量不够充足,影响了模型的准确率[2];第二,以往的研究表明,民众对于房地产市场的预计期望是决定房地产市场价格的重要因素[3],而现有的价格预测中少有考虑民众预期这一指标;第三,现有研究大多使用房价作为房地产市场热度的度量指标,但是由于房地产开发周期较长,加之土地供给量对其价格的约束较大,且与银行信贷关系密切,因此用房价作为整个房地产市场的度量指标有一定的局限性[4]。

住宅销售价格指数是综合反映住宅商品价格水平总体变化趋势和变化幅度的相对数。它通过百分数的形式来反应房价在不同时期的涨跌幅度。其优点是同质可比,这种方法反映的是排除房屋质量、建筑结构、地理位置、销售结构因素影响之后,由供求关系及成本波动等因素带来的价格波动。北京市统计局给出了房价指数的具体计算说明[5],具体操作是在住宅价格的基础上,依据同质可比的原则,按月调查每处住宅的价格变动,对其涨跌幅度进行加权平均,最终得到全市住宅价格的变动幅

收稿日期:2019-08-10
基金项目:河北省住房和城乡建设厅信息中心项目"住房城乡建设行业大数据应用指南"
作者简介:朴春慧(1964-),女,教授,博士,研究方向:大数据处理与分析。
本文信息:朴春慧,武旭晨,蒋学红,等.基于多源数据的房地产价格指数预测模型及实证研究[J].石家庄铁道大学学报:社会科学版,2019,14(2):14-22.

度。假设某市 3、4 月份住宅交易情况如表 1 所示，　　房价指数编制原理可以简化为以下计算过程：

表 1　某市 3、4 月份住宅交易情况

住宅项目	3 月份			4 月份			价格
	建筑面积/m^2	交易金额/万元	单价/(万元·m^{-2})	建筑面积/m^2	交易金额/万元	单价/(万元·m^{-2})	涨幅/%
项目 A	80	320	4.0	80	328	4.1	2.5
项目 B	70	420	6.0	70	434	6.2	3.3
项目 C	135	1 350	10.0	135	1 350	10.0	0.0
项目 D	200	800	4.0	200	860	4.3	7.5

按交易面积加权计算的环比价格指数为：

$$\left(1+\frac{80\times 2.5\%+70\times 3.3\%+135\times 0\%+200\times 7.5\%}{80+70+135+200}\right)\times 100\%=104.0\%$$

按交易金额加权计算的环比价格指数为：

$$\left(1+\frac{328\times 2.5\%+434\times 3.3\%+1\ 350\times 0\%+860\times 7.5\%}{328+434+1\ 350+960}\right)\times 100\%=102.9\%$$

当月价格环比指数为上述环比指数的算术平均数 103.5%。由此可知，房价指数剔除了住宅之间的品质差异，能够更加准确反映全市住宅价格的总体变化程度[5]。

住宅销售价格指数分为城镇新建住宅销售价格指数和二手住宅销售价格指数两部分。其中，城镇新建住宅销售价格指数的统计范围是所有进入房地产市场第一次进行产权交易及网上签约的住宅交易价格，分为保障性住房和新建商品住宅两部分[5]。本文主要研究的是影响因素对于新建商品住宅的非线性映射关系，从而预测其发展趋势，因此本文将国家统计局公布的新建商品房住宅销售价格指数作为房地产价格的度量指标。

针对房地产价格与其影响因素之间复杂的非线性关系，本文使用了两种常用的机器学习算法来预测房地产价格的发展趋势，旨在有效降低由于评估人员主观因素所造成的评估结果的偏差。在实证研究中参照住房和城乡建设部门的政务数据，收集整合了房产市场供求数据、宏观经济调控政策、人们对当前房价的预期和当前本市房产经济发展情况等多源异构数据，建立了两种房地产价格预测模型，并以华北某城市的月度数据为基础，结合 ARIMA 模型和经典 BP 神经网络模型对两种房地产价格评估模型进行了对比分析。

一、房地产价格评估指标体系

房地产市场与社会经济联系密切，同时受国家经济政策和预期计划影响，也与民众对当前房地产价格的预期紧密相关[3]。莫连光[6]使用经济、行政、区域等因素来估算房地产市场价格；王筱欣[7]使用供给因素、需求因素以及经济发展因素对重庆市房价进行了验证与预测。刘佼[1]以成都市为例引入了国民经济和房地产内部协调等指标组成了房地产市场警兆指标体系。本文在全面参考房地产价格评估研究成果和数据挖掘模型性能的基础之上，结合实验数据结构，参照政府部门数据，选定了一种房地产价格评估指标体系。以房地产供求关系、社会宏观经济指标、国家货币政策、民众对房价的预期和房地产价格现状作为一级指标，共 17 项二级指标组成房地产价格评估指标体系（表 2）。为保证预测模型的超前性，本文使用上一个月的指标数据来预测当前月的房地产价格指数。

表 2　房地产价格评估指标体系

一级指标	二级指标
房地产供求关系	商品房销售成交面积/商品房批准预售面积(月)
	住宅实际成交面积/商品房实际成交面积(月)
	新开工面积增长率(%)
	实际成交面积增长率(%)
	二手住宅销售价格指数(上年=100)
社会宏观经济指标	市固定资产投资_累计增长(%)
	市居民消费价格指数(上年同期=100)(%)
	市规模以上工业增加值_累计值(亿元)
国家货币政策	存款基准利率(调整后)(%)
	贷款基准利率(调整后)(%)
	货币供应量(亿元)
	货币供应量同比增长
	大型金融机构存款准备金率(调整后)(%)
	中小型金融机构存款准备金率(调整后)(%)
民众预期	预期下阶段房价上涨的人数占比(%)
房地产价格状况	月新建商品住宅销售价格指数
	月商品房均价(元)

表 2 中，市商品房销售成交面积/商品房批准

预售面积(月)、月新开工面积增长率(%)、月实际成交面积增长率(%)、住宅实际成交面积/商品房实际成交面积(月)、商品房均价(元)来自住房和城乡建设部门信息中心大数据分析平台;市固定资产投资_累计增长(%)来自市统计局大数据平台;二手住宅销售价格指数(上年=100)、城市居民消费价格指数(上年同期=100)(%)、新建商品住宅销售价格指数(上年=100)来自国家统计局;市月度GDP是影响房地产价格的重要影响因素,但当前发布的GDP为季度数据,月度数据缺失,但市规模以上工业增加值_累计值(亿元)月度数据与市月度累计GDP关联度较大,是GDP统计中的重要参考,本文使用市规模以上工业增加值_累计值(亿元)近似表示当月GDP;存款基准利率(调整后)(%)、贷款基准利率(调整后)(%)、大型金融机构存款准备金率(调整后)(%)、中小型金融机构存款准备金率(调整后)(%)、货币供应量(亿元)、货币供应量同比增长来自中国人民银行官方网站;上阶段预期本阶段房价上涨的人数占比(%)来自中国人民银行储户问卷调查报告。

二、BP-Adaboost算法

BP神经网络(Back Propagation Neural Network)是一种使用误差逆向传播算法训练得到的多层前馈型神经网络。其对应算法为BP算法(误差反向传播算法),是一种经典的监督学习算法,其优化目标是使所有样本经过计算后的输出结果与目标输出之间的均方误差最小。算法主要分为两个阶段:信息前馈传递阶段和误差反向传播阶段[8-9]。在信息前馈传递阶段,每层的输入信息首先通过连接权值进行计算,通过相应的激活函数进行变换得到输出信号,再将输出信号作为输入传入下一层继续进行信息变换,最终得到网络输出;在误差反向传播阶段,计算神经网络的输出与真实标签间的误差,通过连接权值从输出层反向传播至输入层,最后依据梯度值更新连接权值。信息前馈传递阶段和误差反向传播阶段构成了一个迭代过程,循环不断地更新神经网络中的权值和阈值,达到预先设置的迭代终止条件后结束,最终实现神经网络中权值和阈值的最优。

Boosting方法是一种用来提高弱分类算法准确度的算法,基本思想是不断使用基础分类模型对数据进行分析建模,在建模过程中通过不断改变错分样品的权重,建立一系列基础分类模型,最后对其进行线性加权组合得到一个强分类器[10]。1995年,Freund and Schapire提出的Adaboost算法是Boosting算法的一个典型代表[11]。其主要流程为:首先给出弱学习算法和样本空间(x, y),从样本空间中找出n组训练数据,每组训练数据的权重都为$1/n$。然后用弱学习算法迭代运算K次,每次运算后按照分类结果更新训练数据的权重分布,对于分类失败的训练个体赋予较大权重,下一次迭代运算时更加关注这些训练个体。弱学习算法通过反复迭代得到一个分类函数序列$f_1, f_2, \cdots, f_K$,每个分类函数赋予一个权重,分类结果越好的函数,其对应权重越大。K次迭代之后,由弱分类函数加权得到最终的强分类函数F[12-13]。

本文使用BP-Adaboost预测器作为房地产价格指数的预测模型之一。BP-Adaboost预测器是以BP神经网络作为模型的弱预测器,通过Adaboost算法得到的由多个BP神经网络组成的一种强预测器。

三、支持向量回归算法

支持向量机(Support Vector Machine, SVM)[14-15]是在统计学习理论的VC维理论和结构风险最小原理的基础上发展起来的一种机器学习方法。支持向量回归是支持向量机在回归问题上的扩展,Vapnik在ε不敏感损失函数的基础上提出了ε支持向量回归机(ε-SVR),它要解决一个原始优化问题:

$$\min\left\{\frac{1}{2}\|w\|^2+\frac{1}{2}\sum_{i=1}^{l}\varepsilon_i+\varepsilon_i^*\right\}$$

$$\text{s.t.}\ [(w\cdot x_i)+b]-y_i\leqslant\varepsilon+\xi_i^*,\ i=1,2,\cdots,l \quad (1)$$

$$y_i-[(w\cdot x_i)+b]\leqslant\varepsilon+\xi_i^*,\ i=1,2,\cdots,l$$

$$\xi_i^*\geqslant 0,\ i=1,2,\cdots,l$$

对于非线性回归问题,引入变换φ,将样本映射到高维空间,再引入Lagrange函数,将凸二次规划问题转化为下面的对偶问题[16]:

$$\min_{\alpha^{(*)}\in R^{2l}}\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(u_i,u_j)+$$

$$\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{l}v_i(\alpha_i^*-\alpha_i) \quad (2)$$

$$\text{s.t.}\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0$$

$$0\leqslant\alpha_i^{(*)}\leqslant C,\ i=1,2,\cdots,l$$

回归估计模型转化为：

$$f(x)\sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(u_i,u_j)+b \qquad (3)$$

式中，$K(u_i,u_j)$为核函数；C 为惩罚参数；$\alpha^{(*)}=(\alpha_1,\alpha_1^*,\cdots,\alpha_l,\alpha_l^*)^T$ 为 Langrangec 乘子向量，α_i 和 α_i^* 为向量中的元素。

本文以支持向量回归机作为房地产市场价格指数预测的模型之一，通过之后的调参工作确定相对参数，使得模型的预测性能最优。

四、实证研究

（一）样本数据的收集与处理

本文以华北某市的房地产市场为示例对象，验证房地产价格指数预测模型的优劣。本文收集了 2010 年 1 月开始至 2017 年 6 月关于此市的各类月度指标数据 90 条（如表 3 所示）。数据集中，固定资产投资、规模以上工业增加值缺失一月份数据，由于其值为累计数据，本文使用每月的平均增长值得到其一月份的估计值。预期下季房价上涨的人数占比指标数据频率较低，需要将低频数据转换为高频月度数据，本文假设当前月度的房价预期与当前季的预期数据相同。

表 3　华北某市房地产市场评估指标数据

变量	日期						
	2010/1	2010/2	2010/3	…	2017/4	2017/5	2017/6
商品房销售成交面积/商品房批准预售面积(月)	0.34	0.51	1.33	…	1.24	0.05	0.76
新开工面积增长率(%)	-71.74	-60.17	32.57	…	-2.36	47.48	-59.71
实际成交面积增长率(%)	-69.38	-39.51	246.1	…	-7.67	32.49	-11.88
住宅实际成交面积/商品房实际成交面积(月)	0.66	0.91	0.85	…	0.78	0.76	0.82
市固定资产投资_累计增长(%)	26.61	39	39.4	…	7.2	7.2	7.3
市规模以上工业增加值_累计值(亿元)	80.6	150.62	266.42	…	608.9	801.95	1015.7
市居民消费价格指数(上年同期=100)(%)	103.1	103.2	102.9	…	100.9	101.3	101.4
存款基准利率(调整后)(%)	2.25	2.25	2.25	…	1.5	1.5	1.5
贷款基准利率(调整后)(%)	5.31	5.3	15.31	…	4.35	4.35	4.35
大型金融机构存款准备金率(调整后)(%)	15.5	16	16.5	…	16.5	16.5	16.5
中小型金融机构存款准备金率(调整后)(%)	13.5	13.5	13.5	…	13	13	13
货币供应量(亿元)	625 609	636 072	649 947	…	1 596 300	1 601 400	1 631 300
货币供应量同比增长	0.259 8	0.255 2	0.225	…	0.105	0.096	0.094
上期预期本期房价上涨的人数占比(%)	0.49	0.49	0.49	…	0.293	0.293	0.293
二手住宅销售价格指数(上年=100)	101.6	101.2	104.4	…	114	111.7	109.2
本月新建商品房均价(元)	5 085.72	4 676.21	4 481.48	…	10 894.07	10 930.08	10 089.61
本月新建商品住宅销售价格指数(上年=100)	103.3	108.4	109.8	…	118	116.8	116.1
下月新建商品住宅销售价格指数(上年=100)	108.4	109.8	110.3	…	116.8	116.1	113.4

由于各影响因素指标的表现形式不同，个别输入分量差距较大，不能体现各分量的同等地位。且输入过大时，网络容易进入 S 型函数的包河区，导致网络无法收敛[12]。因此在网络计算之前需要对样本数据进行标准化处理，以提高网络的训练速度。结合样本数据的特点及量化标准与房地产价格成正比的特性，本文采用了归一化的标准化方法。

为了准确比较两种预测模型的预测性能，本文按照 8∶1 的比例将数据集随机地划分成训练集和测试集，使用训练集分别训练两种预测模型，使用测试集验证模型的预测性能。为了排除随机划分数据集可能造成的偶然性结果，本文将上述实验进行了 6 次，以 6 次实验的均值作为判断模型预测性能的数据依据。

（二）BP-Adaboost 模型的建立

BP-Adaboost 算法的参数主要分为强预测器

Adaboost的训练误差、训练次数,和弱预测器BP神经网络中训练误差、隐含层数、节点个数和传递函数与训练函数。Adaboost中训练误差的设置不可太小,否则容易出现过拟合,也不能太大,容易出现欠拟合。本文借助Matlab R2016a中的神经网络工具箱,建立了BP-Adaboost预测模型。

经过多次重复实验后,本文设置Adaboost的训练误差为2,训练次数为20。使用3层BP神经网络作为弱预测器,根据训练样本设置输入层节点数14个,输出层1个。根据Hornik公式,隐层节点 $N=[\sqrt{2n+m},2n+m]$(其中,n为输入层节点个数,m为输出层节点个数),设置隐含层节点数为20。设定弱预测器BP神经网络的误差精度为0.000 1,隐含层传递函数采用正切Sigmoid函数tansig(),输出层传递函数采用S型激发函数logsig(),网络训练函数采用tringlm(),设置最大训练次数为1 000次,学习率为0.1,目标误差为0.000 1。

Adaboost组合预测模型具体建模步骤如下:

(1)样本数据权重初始化。首次迭代时设置每个样本数据的权重相等,为 $D_1(k)=1/n(k=1,2,\cdots,n)$。

(2)弱预测器预测。每次迭代前将当前的BP网络权值初始化为0,通过训练集训练n个弱预测器。若某一样本数据预测误差大于设定的阈值,表示产生了较大误差,则将其累计权值相加得到这一弱预测器的权值之和:

$$Error_j = Error_j + D_i \tag{4}$$

式中,$Error_j$代表第j个弱预测器权值累加和;D_i代表超过误差阈值的数据的权值。

(3)更新样本数据权重。若当前BP网络的预测结果对此样本误差较小,未超过阈值,则其权值D_i不变。若超过误差阈值,则权值相对增加:

$$D_{i+1}=1.1\times D_i \tag{5}$$

(4)弱预测器权值计算。根据弱预测器权值累加和$Error_j$计算当前BP网络的权值:

$$a_{ij}=\frac{0.5}{\exp(|Error_j|)} \tag{6}$$

(5)构建强预测器。经过n次迭代后得到强预测器:

$$F=a_t\cdot[f_1,f_2,\cdots,f_n] \tag{7}$$

(三)支持向量回归模型的建立

本文借助Python中的Scikit-Learn模块建立了基于支持向量回归算法的价格指数预测模型。Scikit-Learn模块使用了SVR和$NuSVR$两种回归方式,相对于传统SVR,$NuSVR$增加了一个参数nu来控制支持向量的百分比,在使用时与SVR中的参数ε等价。为获得最优的预测效果,本文比较了两种回归方式的优劣,选择性能最佳的回归方式作为最优预测模型。

核函数是支持向量机的核心,核函数的选择直接影响支持向量回归模型的准确度。本文对常用的$Linear$核函数、径向基核函数、$Sigmoid$核函数和$Poly$核函数做了对比分析。结合两种回归算法,共需要28个参数进行优化。具体需要优化的参数如表4所示。

表4中,需要优化的参数用√表示。核函数参数$degree$对应多项式核函数中的参数d;参数$gamma$分别对应多项式核函数、高斯核函数和$sigmoid$核函数中的参数γ;参数$coef0$分别对应多项式核函数和$sigmoid$核函数中的参数r。

表4 不同核函数中需要优化的参数

核函数 kernel	SVR				$NuSVR$			
	Rbf	$Linear$	$Sigmoid$	$Poly$	Rbf	$Linear$	$Sigmoid$	$Poly$
惩罚系数 C	√	√	√	√	√	√	√	√
支持向量下限 nu					√	√	√	√
损失函数 ε	√	√	√	√				
参数 $degree$				√				√
参数 $gamma$	√		√	√	√		√	√
参数 $coef0$			√		√		√	√

由于支持向量回归模型的参数较多,手工调参工作量较大。为简化调参工作,本文设计了一个调参算法对回归模型进行参数优化。由于本文提出的指标体系为时间序列数据,部分指标数据在一段时间内没有改变(如贷款利率),所以在训练集和测试集的划分时需要考虑这种数据对模型

的影响。我们将全部数据随机地划分了 50 次，其中以 80 条数据作为训练集，10 条数据为测试集对模型进行调参工作。以模型在 50 次不同划分方式下对测试集的拟合优度作为回归模型性能优劣的评判标准。算法流程如图 1 所示。

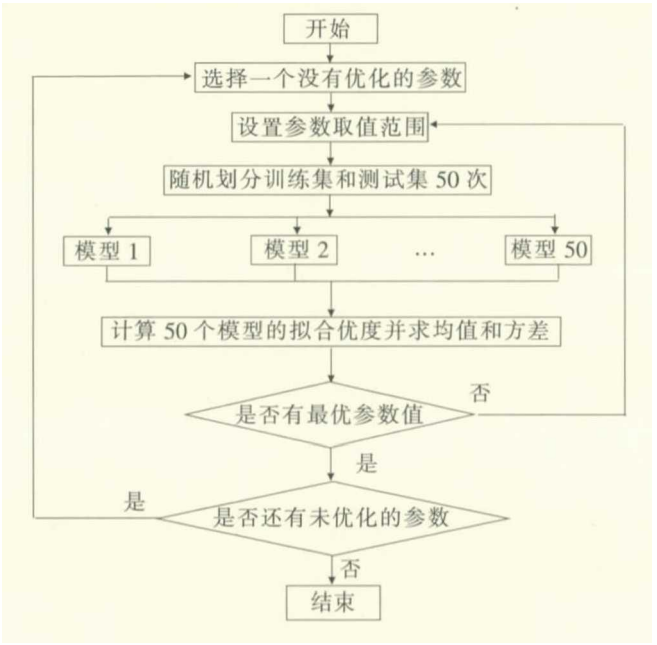

图 1　支持向量回归机参数优化流程图

本文使用决定系数 R^2 对模型的拟合优度进行评价[17]，R^2 越接近于 1 表示模型对数据集的拟合优度越好，其表达式为

$$R^2 = 1 - \frac{\sum (Y_{\text{预测值}} - Y_{\text{实际值}})^2}{\sum (Y_{\text{实际值}} - \overline{Y}_{\text{实际值}})^2} \qquad (8)$$

图 2 给出了 NuSVR-Rbf 模型的参数优化结果，黄色实线表示模型对训练集的拟合优度及其方差，蓝色虚线表示对测试集的拟合优度及其方差。根据图 2 中显示，参数 C 的最优值在 3.0 附近。图 3 给出了不同回归方式下 8 种模型的性能优劣。其中，图 3(a)表示模型对测试集的拟合优度，图 3(b)表示模型对训练集的拟合优度，点为均值，线为方差。实验结果显示，NuSVR-Rbf 模型的性能最优，可作为支持向量回归的最优预测模型。

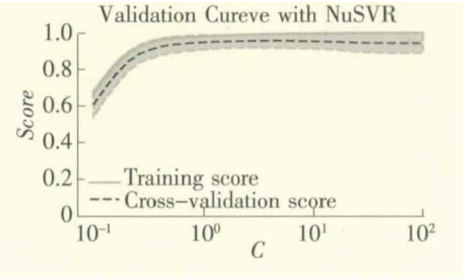

图2　NuSVR-Rbf 中参数 C 的优化示意图

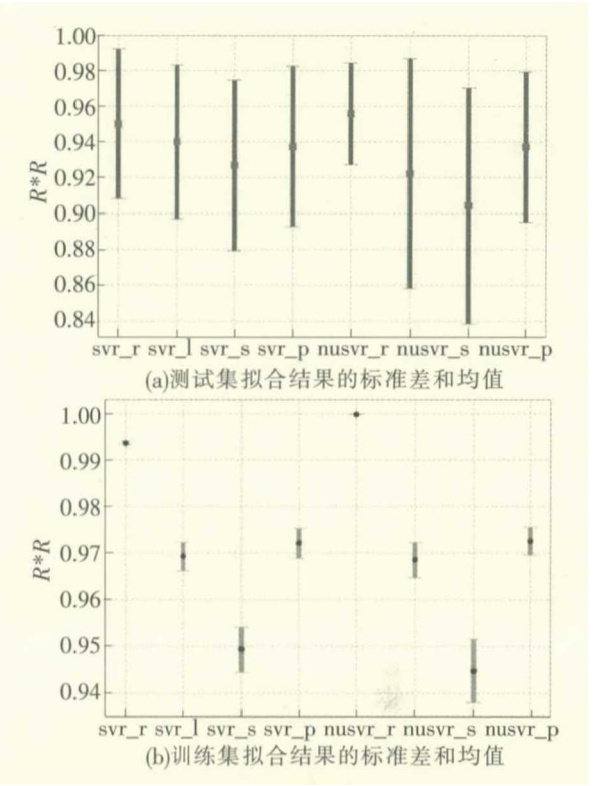

图 3　8 种模型预测性能比较

（四）实验结果及模型比较

为准确判断两种预测模型的预测性能，本文使用 ARIMA 模型和经典 BP 神经网络与本文模型做对比。由于 ARIMA 模型具有不直接考虑其他相关随机变量变化的特点，对未知时间的预测只与时间序列有关，且只能预测未来连续一段时间内价格指数的趋势，而经典 BP 神经网络与本文所提出的支持向量回归模型和 BP_Adaboost 模型可随机预测不同时间的价格指数。为了排除实验过程中随机划分数据集可能造成的偶然性结果，本文按 8∶1 的比例将 90 个月的数据划分为训练集和测试集，并重复划分了 6 次，从而得到不同的数据集 6 组(Dataset1～Dataset6)，以判断经典 BP 神经网络与本文所提模型的预测性能。为判断连续型时间序列作为训练集所得的模型对未来较长时间段的预测精度，本文以数据集中前 80 个月的数据为训练集，后 10 个月的数据为测试集组成数据集 Dateset7，以判断 ARIMA 模型和经典 BP 神经网络模型与本文所提模型在连续型时间序列条件下预测性能的优劣。表 5～表 7 分别展示了 6 组数据集下支持向量回归模型、BP-Adaboost 模型和 BP 神经网络的实验结果，表 8 展示了连续型时间序列下 4 种模型的实验结果。

表5 支持向量回归模型实验结果

Dataset1		Dataset2		Dataset3		Dataset4		Dataset5		Dataset6	
y_pred	y_ture	y_pred	y_ture	y_pred	y_ture	y_pred	y_ture	y_pred	y_ture	y_pred	y_ture
104.276	105.4	109.825	110.1	104.920	103.7	101.216	102.1	100.884	99.8	101.943	101
104.114	104.4	95.081	95.2	105.829	105.3	117.525	116.8	109.758	110.1	101.814	101
101.049	99.6	106.569	105.3	102.420	101.3	102.582	102.1	98.485	97.8	96.114	95.5
106.601	107.8	102.608	103.1	96.578	96.7	110.458	111.6	112.763	111.8	108.713	110.3
117.680	116.8	102.304	103.4	110.234	108.4	109.169	109.9	101.416	101.7	104.364	105.4
96.870	97.5	107.554	107.5	104.075	104.6	107.769	106.7	110.333	109.6	100.017	100
95.256	95.2	105.990	106.3	106.295	106.8	100.580	101.2	102.676	103.1	109.644	109.9
109.507	109.9	107.314	107.8	95.411	95.2	109.793	109.8	98.083	99.6	117.394	119
107.854	108.2	118.083	119	104.309	105.4	107.720	107.8	101.801	101.1	108.175	108.2
99.711	100.6	117.304	118	108.880	109.1	100.570	101.5	117.685	116.8	96.637	97.1

表6 BP-Adaboost 模型实验结果

Dataset1		Dataset2		Dataset3		Dataset4		Dataset5		Dataset6	
y_pred	y_ture	y_pred	y_ture	y_pred	y_ture	y_pred	y_ture	y_pred	y_ture	y_pred	y_ture
104.380	105.4	110.305	110.1	104.982	103.7	101.345	102.1	99.784	99.8	101.841	101
103.913	104.4	95.694	95.2	105.580	105.3	116.789	116.8	110.262	110.1	100.787	101
100.578	99.6	106.592	105.3	102.687	101.3	102.586	102.1	98.875	97.8	95.293	95.5
107.115	107.8	102.496	103.1	96.514	96.7	110.018	111.6	111.436	111.8	110.718	110.3
117.381	116.8	102.748	103.4	108.256	108.4	109.389	109.9	102.136	101.7	104.847	105.4
97.180	97.5	106.767	107.5	103.594	104.6	106.612	106.7	110.365	109.6	99.581	100
95.840	95.2	106.604	106.3	106.156	106.8	99.984	101.2	102.687	103.1	110.700	109.9
109.362	109.9	107.292	107.8	95.588	95.2	110.768	109.8	98.885	99.6	117.391	119
107.180	108.2	120.026	119	104.161	105.4	107.471	107.8	100.786	101.1	107.743	108.2
100.186	100.6	117.298	118	108.754	109.1	101.719	101.5	117.587	116.8	96.784	97.1

表7 BP 神经网络模型实验结果

Dataset1		Dataset2		Dataset3		Dataset4		Dataset5		Dataset6	
y_pred	y_ture	y_pred	y_ture	y_pred	y_ture	y_pred	y_ture	y_pred	y_ture	y_pred	y_ture
106.116	105.4	109.22	110.1	105.219	103.7	102.136	102.1	97.687	99.8	100.781	101
99.378	104.4	97.537	95.2	106.846	105.3	112.407	116.8	109.096	110.1	100.924	101
98.425	99.6	102.005	105.3	99.806	101.3	103.799	102.1	99.397	97.8	97.828	95.5
104.113	107.8	99.366	103.1	95.944	96.7	105.915	111.6	107.758	111.8	96.537	110.3
112.436	116.8	102.993	103.4	104.43	108.4	107.383	109.9	100.766	101.7	100.438	105.4
99.646	97.5	109.937	107.5	101.816	104.6	106.524	106.7	116.616	109.6	96.423	100
97.978	95.2	107.198	106.3	102.769	106.8	102.393	101.2	100.62	103.1	109.246	109.9
111.506	109.9	106.944	107.8	96.773	95.2	114.072	109.8	98.138	99.6	123.679	119
105.845	108.2	118.482	119	93.564	105.4	109.934	107.8	102.781	101.1	106.586	108.2
101.72	100.6	114.135	118	112.068	109.1	102.424	101.5	113.078	116.8	96.515	97.1

表 8　连续时间序列下四种模型的实验结果

变量	日期									
	2017/09	2017/10	2017/11	2016/12	2017/01	2017/02	2017/03	2017/04	2017/05	2017/06
期望值	118.5	118.9	119	118.9	118.5	119	118	116.8	116.1	113.4
ARIMA	119.42	122.1	124.47	126.56	128.4	130.02	131.47	132.75	133.91	134.96
BP-Adaboost	120.11	120.86	120.85	120.19	118.07	117.45	116.6	117.84	116.38	116.51
NuSVR_Rbf	117.02	116.48	115.63	113.81	112.07	112.06	110.86	113.72	112.58	114.05
BP	123.68	127.3	125.75	123.77	125.69	125.09	123.24	123.48	119.63	119.24

考虑到房地产价格指数是一个反映价格变化趋势和变化幅度的相对数，本文使用平均绝对误差 MAE[18] 对模型性能进行评价，如式(9)所示。模型在 6 个随机数据集下的预测精度如表 9 所示。

$$MAE = \frac{\sum | Y_{预测值} - Y_{实际值} |}{n} \tag{9}$$

表 9　四种模型的平均绝对误差

数据集	*NuSVR_Rbf*	*BP-Adaboost*	*BP*	*ARIMA*
Dataset1	0.725 1	0.668 3	2.497	—
Dataset2	0.571 3	0.652	1.923	—
Dataset3	0.737 7	0.690 2	3.248	—
Dataset4	0.667	0.616 5	2.303	—
Dataset5	0.761 8	0.504 7	2.605	—
Dataset6	0.736 1	0.583 3	3.246	—
Dataset7	4.013 9	1.450 5	5.977 8	10.696 1
Average	1.173 3	0.737 9	3.114 3	10.696 1

由表 9 可知，使用随机划分的数据集训练得到的预测模型，其平均绝对误差均比使用连续时间序列数据集训练得到的预测模型小。在所有数据集中，BP-Adaboost 模型对测试集数据预测的平均绝对误差最小。

五、结束语

考虑到房地产市场与其影响因素的非线性映射关系，本文结合房地产供求关系、社会宏观经济指标、国家货币政策、民众对房价的预期和上月房地产价格现状等多源异构数据提出了一套房地产价格评估指标体系。分别使用 BP-Adaboost 算法和支持向量回归算法建立了两个房地产价格指数预测模型，以华北某市为对象对预测模型做了示例研究，并与 ARIMA 模型和经典 BP 神经网络模型作对比。实验结果表明，使用随机划分的数据集训练得到的预测模型比使用连续时间序列数据集训练得到的预测模型误差小。推测可能是由于预测时间段内房地产市场出现了变化，过去的预测模型不再适用。同时，相较于其他三种模型，BP-Adaboost 模型的预测误差最小，使用 BP-Adaboost 模型预测房地产价格指数具有可行性。

参考文献：

[1]刘佼，袁红平. 基于人工神经网络的房地产市场预警模型研究——以成都市为例[J]. 工程管理学报，2016(2)：147-152.

[2]张彦周，马秋香. 基于 BP-Boosting 算法的商品住宅价格预测模型[J]. 河南科学，2014(12)：2588-2592.

[3]薛志勇. 基于预期理论的房地产宏观政策效果的影响分析[D]. 北京：中国科学技术大学，2012.

[4]陈日清. 中国货币政策对房地产市场的非对称效应[J]. 统计研究，2014(6)：33-41.

[5]北京市统计局. 2019-02 住宅销售价格指数[DB/OL]. (2019-3-15). [2019-4-16]. http://tjj.beijing.gov.cn/tjsj/yjdsj/fj_5661/2019/ 201903/t20190315_418798.html.

[6]莫连光. 房地产税开征背景下商业房地产价格估算——基于改进的粒子群算法[J]. 中南财经政法大学学报，2014(4)：38-43，158.

[7]王筱欣，高攀. 基于 BP 神经网络的重庆市房价验证与预测[J]. 重庆理工大学学报：社会科学版，2016，30(09)：49-53.

[8]刘威，刘尚，白润才，等. 互学习神经网络训练方法研究[J]. 计算机学报，2017，40(6)：1291-1308.

[9]吕霁. 基于遗传算法优化神经网络的房地产评估模型及实证研究[J]. 计算机科学，2014，(S2)：75-77，87.

[10]张圆圆，侯艳，李康. 多分类研究中的 boosting 算法[J]. 中国卫生统计，2018，35(1)：142-145.

[11]Wu H，Zou B J，Zhao Y Q，et al. An automatic video text detection method based on BP-adaboost[J]. Multimedia Tools & Applications，2016，75(13)：

1-24.

[12]Cao J, Chen L, Wang M, et al. A Parallel Adaboost-Backpropagation Neural Network for Massive Image Dataset Classifica-tion[J]. Scientific Reports, 2016, 6:38201.

[13]王小川. MATLAB神经网络43个案例分析[M]. 北京:北京航空航天大学出版社, 2013:12-14.

[14]Zhang N, Zhang Y, Wang X. Forecasting of Short-Term Urban Rail Transit Passenger Flow with Support Vector Machine Hybrid Online Model[C]// Transportation Research Board 92nd Annual Meeting,2013.

[15]王振武. 数据挖掘算法原理与实现[M]. 2版. 北京:清华大学出版社, 2017:154-155.

[16]尹强,佐磊,何怡刚,等. 基于支持向量回归机的RFID室内定位研究[J]. 计算机工程与科学,2017, 39(12):2340-2344.

[17]Danuel T,Chantal D. 数据挖掘与预测分析[M]. 2版.王念滨,宋敏,裴大茗,译. 北京:清华大学出版社,2017(2):422-423.

[18]李瑶,曹菡,马晶. 基于改进的灰色模型的旅游需求预测研究[J]. 计算机科学,2018,45(1):122-127.

Prediction Model and Empirical Study of Real Estate Price Index Based on Multi-source Data

Piao Chunhui[1], Wu Xuchen[1,2], Jiang Xuehong[3], Li Yuhong[4]

(1. School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043,China;
2. Ministry of Information Technology, Bank of China Hebei Branch, Shijiazhuang 050000,China;
3. Department of Housing & Urban-Rural Development. Hebei, Shijiazhuang 050051, China;
4. School of Economics and Management, Shijiazhuang Tiedao University, Shijiazhuang 050043,China)

Abstract: The price change of real estate has a significant impact on social and economic development. It is particularly important to accurately predict the price change of real estate market and to effectively control it. However, the use of house price as a measurement index to evaluate the real estate market has certain limitations. Residential sales price index is a relative index issued by the State Statistical Bureau, which comprehensively reflects the general trend and range of changes in housing commodity prices. In order to explore the forecasting methods and effectiveness of the new commercial housing sales price index, it utilizes related real estate supply and demand relationship, social macroeconomic indicators, national monetary policy and people's expectations of housing prices. According to the data, a set of real estate price index system is constructed. Two machine learning algorithms, BP-Adaboost and Support Vector Regression Machine (SVR), are used to construct the real estate evaluation model, and a parameter adjustment algorithm is designed to optimize the parameters of the SVR model. The monthly real estate data of a city in North China are used to train and forecast the two models, and The ARIMA model and the classical BP neural network model are compared with the model proposed in this paper. The experimental results show that the prediction error of BP-Adaboost model is the smallest, and it is feasible to use BP-Adaboost model to predict the real estate price index.

Key words: real estate forecasting; BP-adaboost algorithm; support vector regression; residential sales price index