文章编号：1007-757X(2021)06-0042-04

基于 RNN 的房地产估价回归模型

谢志伟

（东莞职业技术学院 计算机工程系，广东 东莞 523808）

摘　要：随着机器学习及相关领域的飞速发展，机器学习也随之进入到房产行业，科学地对房价进行评估和预测。由于现有的方法存在着精度不高等问题，提出一个新的更复杂的分析模型。该模型是基于所收集的房屋数据信息，然后利用一个递归神经网络结合 XGBoost 树的新模型，对这些数据进行分析，从而实现对房价的预测。通过实验表明，这个模型相对于现有的模型，误差减小了近 15%。因此能够满足实际需求。

关键词：机器学习；房价估计；回归模型；XGBoost；RNN；LSTM

中图分类号：F293.3；TP391.3　　**文献标志码：**A

Regression Model of Real Estate Valuation Based on RNN

XIE Zhiwei

(Department of Computer Engineering, Dongguan Polytechnic, Dongguan 523808, China)

Abstract: With the rapid development of machine learning and related fields, machine learning has also entered the real estate industry to scientifically evaluate and predict housing price. Because of the problem of low precision in existing methods, a new and more complex analysis model is proposed to solve the above shortcomings. The model is based on the information collected from housing data, and then uses a recurrent neural network combined with the new model of XGBoost tree to analyze the data to achieve the prediction of housing price. Experiments show that the model of this study reduces the error by nearly 15% compared with the existing model, so it can meet the actual needs.

Key words: machine learning; house assessment; regression model; XGBoost; RNN; LSTM

0　引言

对大多数人来说，住房一直是最大的开支之一。买房是一个高度参与的决定。消费者对房产价值的判断和对房产未来价值的估计，会影响他们的购买决策和预算分配[1]。此外，房地产价格是反映经济活动的重要因素之一。因此，对土地价格的准确预测，可以帮助政府或企业在未来的财政年度内做出操纵财务状况的关键决策。从这个角度看，房地产价格的测算过程与人们的生活和国民经济息息相关[2]。

自动估价模型（AVM）是在分析房地产的区位、周围条件和特性的基础上，对房地产市场价值进行评估的数学程序[3]。房地产行业的一些企业提供了易于访问的 AVM Web 应用程序来估计房产价格，主要是基于套索回归（LASSO）和支持向量回归（SVR）[4-5]。但这些方法没有更多考虑房屋本身的属性，如房间数量、房屋大小和房屋的装修情况等。所以为了更加准确地评估房产价格，在此基于递归神经网络（RNN）和房屋自身属性，提出了一种新的房地产价格评估方法。同时，Boosting 树模型作为数据分析竞争中一种很有前途的机器学习方法[6-7]。因此，在本研究中，为了使结果更加准确，通过 RNN 网络模型与 Boosting 树的一种变体，XGBoost 模型相结合，对房价进行预测。

1　基于 LSTM 和 XGBoost 的模型

在本节中，简要介绍所提出模型的主要组成部分。首先，介绍 RNN 中的长期短期记忆（LSTM）的基本体系结构，然后介绍了 XGBoost 模型。

1.1　长期短期记忆

在自然语言处理（NLP）中，整个句子被定义为顺序数据，每个词都基于对先前词的理解。当人工神经网络执行自然语言处理时，它需要一种结构来根据句子的上下文来推理下一个单词，该结构将先前的输出作为推论的输入进行组合。递归神经网络（RNN）是用于处理顺序数据的一系列神经网络[8-9]。

RNN 结构示意图如图 1 所示。

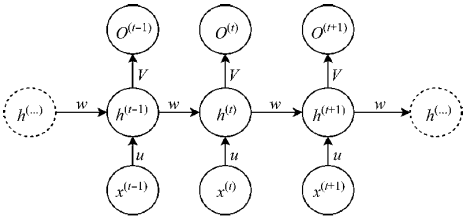


图 1　RNN 结构示意图

图 1 说明了简单 RNN 的结构。$\{O^{(1)},\cdots,O^{(T)}\}$ 是给定

基金项目：广东省科技计划项目（2016B050502001）；东莞职业技术学院政校行企合作开展科研与服务项目（政 201817）

作者简介：谢志伟（1979-），男，硕士，副教授，研究方向：系统开发与设计等。

输入序列$\{x^{(1)},\cdots,x^{(T)}\}$和隐藏单元的神经网络的隐藏层$\{h^{(1)},\cdots,h^{(T)}\}$。来自输入单元的单向信息流到达隐藏单元，而来自隐藏单元的另一单向信息流到达输出单元。$h^{(t)}$是基于当前输入层的输出和先前隐藏层$h^{(t-1)}$的状态来计算的，估算方法如式(1)。

$$h^{(t)} = f(Ux^{(t)} + Wh^{(t-1)}) \tag{1}$$

式中，f表示非线性激活函数，如 tan 或 ReLU，具有共享参数U,W。$O^{(t)}$是步骤t的输出，它取决于当前神经元的激活函数，如式(2)。

$$O^{(t)} = \sigma(Vh^{(t)}) \tag{2}$$

式中，σ表示输出层的激活函数。

从理论上讲，RNN 可以从句子开始处理上下文，这样可以更准确地预测句子结尾的单词。然而，序列长度越长，隐藏层就越多，这就产生了消失梯度问题，从而阻碍了 RNN 的优化[8]。

LSTM 是解决这个问题的架构[10]，每个 LSTM 将整个神经网络分割成多个单元$\{C^{(1)},\cdots,C^{(T)}\}$，如图 2 所示。

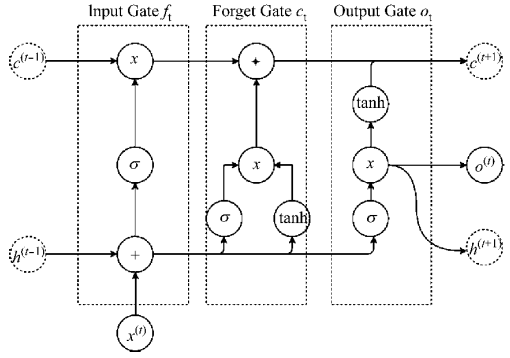


图 2　LSTM 的结构

每个单元包含输入门、遗忘门和输出门，其能够在正向传播阶段存储错误。遗忘门将误差从单元中删除，以求解消失梯度。

W_f、W_c和W_o分别是输入门、遗忘门和输出门的对应参数。输入门将电流输入和先前的输出结合起来，在神经元中使用激活函数σ和偏置b_f。然后，tan 为单元值创建新的候选值，并分别用偏差b_i和b_c与先前的更新决策值进行比较，如式(3)—式(5)。

$$f_t = \sigma(W_f[h^{(t-1)},x^{(t)}] + b_f) \tag{3}$$

$$c_t = \sigma(W_c[h^{(t-1)},x^{(t)}] + b_i) * \tan(W_c[h^{(t-1)},x^{(t)}] + b_c) \tag{4}$$

$$o_t = \sigma(W_c[h^{(t-1)},x^{(t)}] + b_o) * \tan(c_t + f_t) \tag{5}$$

1.2　XGBoost 原理

XGBoost 是 Boost 算法的一种，是基于 gradientboosting 框架实现的[11-12]。它是一个分布式梯度的优化增强库，由很多分类回归树组成。由于 XGBoost 可以进行多线程计算，所以它具有运算速度快、体积小的特点[13-14]。XGBoost 算法核心是为了拟合前一次迭代中实际值和预测值的差，所以在每次迭代的过程中都会增加一棵树，从而让预测值不断接近真实值。然后每棵树的总得分就是该样本的得分。XGBoost 的预测值计算如式(6)。

$$\hat{y}_j^{(g)} = \sum_{k=1}^{g} f_k(s_r) = \hat{y}_j^{(g-1)} + f_g(s_r),\quad f_g \in F, r \in n \tag{6}$$

式中，$\hat{y}_j^{(g)}$表示预测值；n表示总样本个数；r表示第r个样本；g表示决策树个数；f表示第g个决策树；j表示第j个样本；F表示集合空间。得到损失函数的结果如式(7)。

$$L^{(g)} = \sum_{r=1}^{n} d_r + \sum_{k=1}^{g} \Omega(f_k) \tag{7}$$

式中，d_r表示预测值和实际值的偏差程度。$\Omega(f_k)=\gamma^T+\frac{1}{2}\lambda\|\omega\|^2$，在这里，$T$是子节点的个数，$\lambda$和$\gamma$是正则化因子。因此，最优目标函数为式(8)。

$$L^{(g)} = -\frac{1}{2}\sum_{m=1}^{T}\frac{P_m^2}{V_m+\lambda} + \gamma T + U \tag{8}$$

式中，P表示损失系数；C表示损失因子；V表示分裂的节点数。在 XGBoost 中判断节点是否进行分裂的方法是通过分裂后的左右节点的分数减去未分裂的节点分数。由于 XGBoost 中利用正则化因子来限制树的增长，所以当收益小于正则化因子时，节点分裂则停止。整个 XGBoost 的流程如图 3 所示。

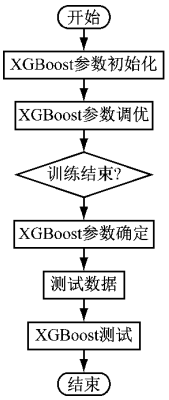


图 3　XGBoost 预测流程图

2　模型建立

一般来说，图像包含的有价值的信息不能简单地写下来，例如，属性的质量或状态是什么？它看起来如何？这些颜色是否很好地融合并增强了房屋的外观和感觉？所以在此，希望有一个图像评估模型可以给定一个图像作为输入，自动分配一个评分，可以模仿人类来观察和欣赏其价值，并从不同的图像中评估房屋属性。

2.1　数据预处理

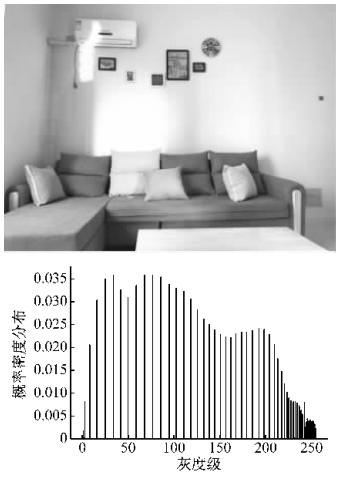
由于相关房屋照片的尺寸大小不一，所以在进行特征提取之前，需要对图片数据进行预处理。首先先要将图片尺寸统一，在此，LSTM 网络的输入尺寸是 244×244 像素的图像。同时，由于提出的 LSTM 网络需要对房屋多个属性进行评分，因此，在此将输入的图像切割成 122×122 像素的 4 个小图像。

由 4 个小图像构成整个大的输入图像，如图 4 所示。

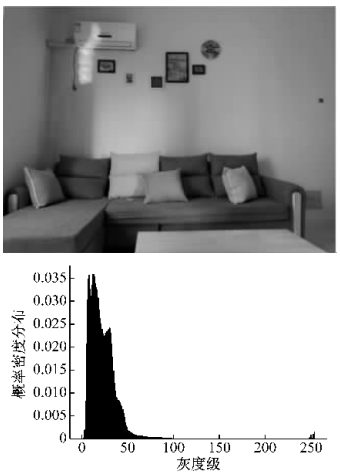
同时，由于房屋图片存在通过调亮光线进行美化的情

图 4　LSTM 输入图像示例

况，因此，为了使得整个模型对房屋的评估更加准确，所以对于美化过的房屋图片需要进行反美化处理。为了确定图像是否经过光线增强处理，首先需要统计同一房屋其他区域的图片及相似房源的图片的光线强度。因此，将原 RGB 图像转为 YCbCr 图像，然后计算每一幅图像的光亮值，对参考图像的光亮值进行平均处理，确定阈值。如果高于阈值则说明该幅图像经过美化处理，需要调低光亮，如图 5 所示。



(a) 未处理的图像及其均衡化直方图



(b) 反增强后的图像及其直方图

图 5　去美化前后图片对比

该图显示了去美化前后图像对比。

2.2　特征的选择和提取

尽管网上的房产图片可以对一栋房子进行整体评价，但却不能捕捉到一些特征，如窗户、门、镜子、屋角等。文献[15]指出，从房地产图像中提取视觉特征与正常属性有显著关系，可以提高房价估计的准确性。因此，利用 RNN 神经网络中的 LSTM 网络进行图像的特征提取和视觉特征学习。

输入层是对应的视频帧特征向量，在输入层上层是正向的 LSTM 层，由一系列的 LSTM 单元构成。再将全部时刻的 LSTM 输出进行加权平均操作后的结果作为上层的表示。最后通过 softmax 层，进行全连接的操作。

数据集中有太多用于建模的变量，选择这些功能有两个原因。一是特征集过大会使算法速度变慢；二是当变量的个数明显高于最优值时，会导致机器学习的不精确性。因此，根据真实性和相关性来选择最佳特征是至关重要的。Boruta 是一种基于随机森林的特征选择方法，并应用于我们的实验中进行特征提取。在特征选择之后，只有部分特征被用来构建模型。特征选择的结果包括有楼房单元号、屋顶类型、房间数、附加设施和地址等一系列与房产有关的因素。

每个特征的相关属性都有不同数量的图像，其中有些属性有 5 个图像，有些属性有大约 35 个图像。通过对现有的数据进行统计，大部分记录都有 10 到 30 幅房产图片。对于构建此模块，将删除少于 10 个图像或多于 30 个图像的属性记录。受文献[16]发表的神经图像评估的启发，属性平均质量评分可以定义为式(9)。

$$\mu = \frac{1}{M}\sum_{t=1}^{M}\left(\sum_{i=1}^{N} S_i \times P_i\right)_t \tag{9}$$

式中，M 表示每个属性的图像总数，对于这个实验，M 被设置为 $5 \leqslant M \leqslant 30$，因为大多数属性都在这个范围内；$S$ 表示 1 到 10 的评分等级，所以 S 的范围为 $S \in [1,10]$；N 表示总分列数，通过大样本分析，N 设置为 10，这意味着它有 10 列评分；P 表示每个评分的响应百分比。

2.3　价格预测

这一部分说明了房价预测模型的具体流程，该模型结合了一些用于房价预测的特征。混合模型包括在数据集上预先训练的 LSTM 模型，具有 softmax 功能，用于评估房产图像，并给出总体房屋评分；激活校正线性单位(ReLU)以分析表格数据集/数字特征；另一个具有 ReLU 激活功能的 LSTM 模型用于从属性图像中提取视觉特征，作为属性评估的附加属性；用 XGBoost 预测房地产价格。

3　实验与评估

3.1　实验环境与数据

本文的实验环境是基于一台联想 ThinkPad 笔记本电脑，其处理器为英特尔 I7 处理器，显卡为英伟达 Quadro T2 000，内存大小为 16GB，系统为 windows 10 64 位系统。

在整个实验中，数据都是来自于 Data Nerds 的数据库。收集的数据来自美国最大城市之一的伊利诺伊州的芝加哥市，以及美国房产的多重上市服务系统中的图片数据。本节介绍如何与 SVR 和 LASSO 回归相比，对数据进行预处理和评估所提出的模型。整个数据集随机分成抽取 80%的数据

作为训练集，剩下 20%的数据作为测试集。

3.2　数据集预处理

美国房价指数(Housing Price Index，HPI)数据集由联邦政府提供。整个数据集包含 1979 年至 2019 年美国所有地级市的所有 HPI。在这个实验中，我们提取了芝加哥邮政编码级别的 60 个 HPI 系列。

原始数据集包含许多变量，如房屋质量，房产地理信息。它还包含了房价随时间变化的交易记录。在这里，只选择了 2017 年内，并通过 HPI 将 2018 年和 2019 年的价格转换为该实验的真实数据。在全市范围内筛掉了价格极高或极低的房子，筛选数据的摘要如表 1 所示。

表 1　芝加哥的平均价格和标准价格偏差

	房屋总数	平均价格	标准价格偏差
芝加哥(Chicago)	5 179	208.2	118.6

为了训练和验证提出的模型并防止过度拟合，采用了 5 倍交叉验证技术。该算法将完整的数据随机分成五个子集。一个唯一的子集作为测试的验证数据，其余四个子集用于每个验证过程中的训练。经过 5 倍交叉验证，我们可以得到每套房子的预测价格。

3.3　训练方法

模型训练过程，如图 6 所示。

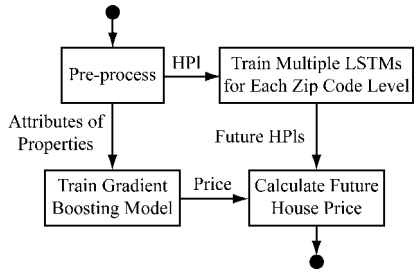


图 6　模型训练过程

首先，预处理后的过滤数据包含 1979 年至 2017 年的房屋和其属性，如前所述。其次，采用多个 LSTMs 分别对每个邮政编码级别的 HPIs，以及房屋自身照片进行评分和预测。它是一个具有 4 个激活 ReLu 神经元的单隐层 LSTM，窗口大小是 3，这意味着预测 HPI 是由前 3 个 HPI 预测的。同时，XGBoost 模型有义务根据房产属性预测 2017 年的房价。最后，利用预测的 2017 年的结果对 2018 年和 2019 年的房价进行评估。

3.4　评估模型和实验结果

模型对于不同房屋的评分结果如图 7 所示。

由图 7 可知，两个房屋的评分均显示在卧室图片中，可以看到(a)图的评分高于(b)图，这与实际结果也是相同的。

在整个评估过程中，所采用的评价指标为平均绝对误差(MAE)和平均绝对百分比误差(MAPE)。两个度量的定义，如式(12)、式(13)。

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|\mathrm{ture}_i-\mathrm{pred}_i\right| \tag{12}$$

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\mathrm{ture}_i-\mathrm{pred}_i}{\mathrm{pred}_i}\right| \tag{13}$$

式中，ture_i 表示真实值；pred_i 表示预测值。

(a) A 房屋卧室图

(b) B 房屋卧室图

图 7　模型评分结果

在此使用相同的训练和测试集来评估所有的模型。所有不同模型的回归结果如表 2 所示。

表 2　结果比较

	MAE	MAPE
LASSO	58.67	31.35%
SVR	57.23	29.94%
提出的方法	46.91	24.02%

结果表明，提出的方法，相对于 LASSO 回归，误差减小了近 15%，相对于 SVR 回归，误差减小了 10%，所以该模型比其他两个模型具有更好的性能。

4　总结

本文提出了一种用于房地产估价的集成学习回归模型。该模型能够综合考虑房屋质量、区位和市场价格走势。实验结果表明了所提出方法是有效的，也为深度学习方法与统计学习算法的集成提供了一种新的途径。这也说明了深度学习在房地产领域具有广阔的未来。

参考文献

[1] 江华波. 房地产项目投资决策要点分析及建议[J]. 中国集体经济，2020(4)：61-62.

[2] 马荣华. 房地产经济对中国国民经济增长的作用影响研究[J]. 现代经济信息，2018(21)：2-3.

[3] 陈敏，李英冰. 基于特征价格理论和神经网络的武汉二手房价自动评估[J]. 城市勘测，2018(4)：21-24.

[4] 武杰. Logistic 回归中的随机 Lasso 方法[D]. 北京：北京建筑大学，2018.

(下转第 54 页)

表 1　对照组和实验组成绩显著性差异分析结果

(a) 分组统计数据

	分组	个案数	平均值	标准差	标准误差平均值
成绩	实验组	31	66.81	7.490	1.345
	对照组	31	61.03	5.148	.925

(b) 独立样本 T 检验

		方差相等的 Levene 检验		等均值假设的 t-检验						
									差异的 95%的置信区间	
		F 检验	显著性	t 值	自由度	显著性(双侧)	均值差值	标准误差值	下限	上限
成绩	假设方差相等	5.605	.021	3.538	60	.001	5.774	1.632	2.509	9.039
	假设方差不相等			3.538	53.173	.001	5.774	1.632	2.501	9.048

能够有效地促进学生学习的自主性和能动性，为学生构建有意义的知识体系创建了协作、探究式的学习情境，最终显著地提高了高级英语学习成绩。

4.2　访谈结果及讨论

为了调查学生对混合式教学模式的满意度，从实验组随机抽选了 22 人进行访谈。访谈内容由 3 个开放式问题构成：1. 你对线上＋线下的混合式教学模式满意吗？2. 你认为混合式教学模式能够促进学生自主性学习吗？3. 你认为混合式教学模式能够提高高级英语学习成绩和学习兴趣吗？基于上述 3 个问题，回收有效访谈 22 份。数据结果显示 82%的学生对线上＋线下的混合式教学模式满意，89%的学生认为混合式教学模式能够促进学生的自主性学习，80%的学生认为混合式教学模式能够提高其学习成绩和学习兴趣。上述访谈数据表明学生对混合式教学模式在高级英语课程中的应用效果持肯定态度，这一结果与测试数据结果相吻合，也支持了建构主义学习观点。

5　总结

上述研究结果表明，混合式教学模式能够有效提高学生成绩，并收到学生满意度较高的评价。混合式教学模式把线上网络资源平台资源与线下面对面课堂教学有机结合，以任务为驱动，为学生自主性学习创建了有利情境，促进了学生与网络平台资源、学生与学生、学生与教师的高质量互动协作，培养学生独立探索、思考的能力，使得学生逐渐形成批判性思维能力。这为其他课程进行混合式教学模式改革研究提供了借鉴意义。

参考文献

[1] 何克抗. E-Learning 的本质——信息技术与学科课程的整合[J]. 电化教育研究，2002，23(1)：3-6.

[2] 郭欣宇，曹锦华，刘阳. 基于微信和微课的高级英语翻转课堂教学研究[J]. 大庆师范学院学报，2019，39(6)：120-128.

[3] 范丽娟. 交往理论视域下《高级英语》"学习共同体"教学模式研究[J]. 黑龙江高教研究，2020，38(1)：157-160.

(收稿日期：2020.09.10)

(上接第 45 页)

[5] 陆鹏. 上海商品房价格走势预测方案设计：基于文本挖掘与支持向量机[D]. 上海：上海师范大学，2018.

[6] 张先勇，汤鲲. 基于 XGBoost 算法结合域名信息筛选的流量识别方法[J]. 电子设计工程，2019，27(6)：177-182.

[7] 江凯，王守东，胡永静，等. 基于 Boosting Tree 算法的测井岩性识别模型[J]. 测井技术，2018，42(4)：395-400.

[8] Al-Smadi M, Qawasmeh O, Al-Ayyoub M, *et al*. Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels' reviews[J]. Journal of Computational Science, 2018, 27(7):386-393.

[9] 高茂庭，徐彬源. 基于循环神经网络的推荐算法[J]. 计算机工程，2019，45(8)：198-202.

[10] 姚开一，李英玉. 基于神经网络的地震震相自动拾取方法[J]. 电子设计工程，2018，26(22)：1-5.

[11] 李曼洁，吴照，徐斌辰，等. 汽车制造厂油漆车间能耗预测模型[J]. 微型电脑应用，2019，35(3)：1-4.

[12] 沈夏炯，张俊涛，韩道军. 基于梯度提升回归树的短时交通流预测模型[J]. 计算机科学，2018，45(6)：222-227.

[13] 甘鹭. 基于机器学习算法的信用风险预测模型研究：以互联网金融公司数据样本为例[D]. 北京：北京交通大学，2017.

[14] Zhang D H, Qian L Y, Mao B J, et al. A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGBoost[J]. IEEE Access, 2018, 6:21020-21031.

[15] Ahmed E H, Moustafa M. House Price Estimation from Visual and Textual features[C]// Proceedings of the 8th International Joint Conference on Computational Intelligence. November 9-11, 2016. Porto, Portugal. SCITEPRESS-Science and Technology Publications, 2016: 1-7.

[16] Talebi H, Milanfar P. NIMA: Neural Image Assessment[J]. IEEE Transactions on Image Processing, 2018, 27(8):3998-4011.

(收稿日期：2020.09.29)