Journal of Computer Applications
计算机应用,2020,40(9):2762－2767
ISSN 1001-9081
CODEN JYIIDU
2020-09-10
http://www.joca.cn

文章编号:1001-9081(2020)09-2762-06
DOI:10.11772/j.issn.1001-9081.2019122249

基于集成LightGBM和贝叶斯优化策略的房价智能评估模型

顾　桐[1,2],许国良[2*],李万林[2],李家浩[1,2],王志愿[2],雒江涛[2]
(1.重庆邮电大学 通信与信息工程学院,重庆 400065;　2.重庆邮电大学 电子信息与网络工程研究院,重庆 400065)
(*通信作者电子邮箱xugl@cqupt.edu.cn)

摘　要:针对传统房价评估方法中存在的数据源单一、过分依赖主观经验、考虑因素理想化等问题,提出一种基于多源数据和集成学习的智能评估方法。首先,从多源数据中构造特征集,并利用Pearson相关系数与序列前向选择法提取最优特征子集;然后,基于构造的特征,以Bagging集成策略作为结合方法集成多个轻量级梯度提升机(LightGBM),并利用贝叶斯优化算法对模型进行优化;最后,将该方法应用于房价评估问题,实现房价的智能评估。在真实的房价数据集上进行的实验表明,相较于支持向量机(SVM)、随机森林等传统模型,引入集成学习和贝叶斯优化的新模型的评估精度提升了3.15%,并且百分误差在10%以内的评估结果占比84.09%。说明所提模型能够很好地应用于房价评估领域,得到的评估结果更准确。

关键词:多源数据;特征选择;轻量级梯度提升机;集成学习;贝叶斯优化;房价智能评估
中图分类号:TP399　　**文献标志码:**A

Intelligent house price evaluation model based on ensemble LightGBM and Bayesian optimization strategy

GU Tong[1,2], XU Guoliang[2*], LI Wanlin[2], LI Jiahao[1,2], WANG Zhiyuan[2], LUO Jiangtao[2]
(1. *College of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*;
2. *Electronic Information and Networking Research Institute, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*)

Abstract: Concerning the problems in traditional house price evaluation method, such as single data source, over-reliance on subjective experience, idealization of considerations, an intelligent evaluation method based on multi-source data and ensemble learning was proposed. First, feature set was constructed from multi-source data, and the optimal feature subset was extracted using Pearson correlation coefficient and sequential forward selection method. Then, with Bagging ensemble strategy used as a combination method, multiple Light Gradient Boosting Machines (LightGBMs) were integrated based on the constructed features, and the model was optimized by using Bayesian optimization algorithm. Finally, this method was applied to the problem of house price evaluation, and the intelligent evaluation of house prices was realized. Experimental results on the real house price dataset show that, compared with traditional models such as Support Vector Machine (SVM) and random forest, the new model introduced with ensemble learning and Bayesian optimization improves the evaluation accuracy by 3.15%, and the evaluation results with percent error within 10% account for 84.09%. It can be seen that, the proposed model can be well applied to the field of intelligent house price evaluation, and has more accurate evaluation results.

Key words: multi-source data; feature selection; Light Gradient Boosting Machine (LightGBM); ensemble learning; Bayesian optimization; intelligent evaluation of house price

0　引言

党的十九大报告中明确提出:要推动互联网、大数据、人工智能与实体经济的深度融合。房地产作为影响实体经济的关键因素,在实体经济发展中发挥着至关重要的作用。伴随着房地产市场化的推进以及市场经济体制的发展,房地产市场对房价评估的需求迅速增长。房价评估有利于房地产市场的决策过程,进而推动经济效益和社会效益的提升。

近年来,国内外学者针对房价评估问题展开了大量研究。例如:Liu等[1]提出了一种基于数据挖掘的双支持向量机模型,用于评估二手房的价格。Phan等[2]提出了基于逐步回归和支持向量机(Support Vector Machine, SVM)相结合的房价

收稿日期:2020-01-09;**修回日期:**2020-02-25;**录用日期:**2020-03-13。
基金项目:教育部-中国移动科研基金资助项目(MCM20170203);重庆市自然科学基金资助项目(cstc2018jcyjAX0587);重庆市技术创新与应用示范(产业类重点研发)项目(cstc2018jszx-cyzdX0124)。
作者简介:顾桐(1995—),男,四川南充人,硕士研究生,主要研究方向:机器学习、数据挖掘;　许国良(1973—),男,浙江金华人,教授,博士,主要研究方向:光电传感与检测、通信网络设计与规划、大数据分析挖掘;　李万林(1963—),男,四川广安人,教授,博士生导师,博士,主要研究方向:新一代网络技术、自动驾驶、车联网、移动大数据;　李家浩(1994—),男,重庆永川人,硕士研究生,主要研究方向:数据挖掘;　王志愿(1995—),男,河南驻马店人,硕士研究生,主要研究方向:数据挖掘;　雒江涛(1971—),男,河南郑州人,教授,博士生导师,博士,主要研究方向:移动大数据、新一代网络技术、通信网络测试与优化。

评估方法。Feng等[3]通过构建多层级模型和人工神经网络(Artificial Neural Network，ANN)的方法对房价进行评估。Mukhlishin等[4]比较了模糊逻辑、ANN和K近邻算法(K-Nearest Neighbor，KNN)在房价评估中的应用。Lu等[5]提出了一种基于Lasso和梯度提升回归的混合模型用于评估房价。王昕睿[6]通过加权求和的方式融合梯度提升决策树(Gradient Boosting Decision Tree，GBDT)、随机森林和反向传播(Back Propagation，BP)神经网络，并以多层级的集成策略实现房价评估。刘燕云[7]分别构建随机森林、SVM、Boosting等单一评估模型，然后采用Stacking算法组合各模型。实验表明，Stacking算法不仅降低了评估的误差，还提升了模型的泛化能力。陈敏等[8]建立了一种神经网络分级模型，用于二手房价格的评估。王海泉[9]通过多元线性回归、神经网络和随机森林分别对房价进行评估。李恒凯等[10]融合地理信息系统(Geographic Information System，GIS)和BP神经网络对房价进行评估，结果表明模型具有较高的精度。

然而，这些房价评估方法采用单一的模型或者经过简单融合的集成模型，泛化性能较差；考虑的因素也不全面，忽略了特征选择和参数组合对模型的影响，模型精度有限。

针对上述问题，本文从多源数据的角度出发，提出一种集成LightGBM(Light Gradient Boosting Machine)模型，并利用贝叶斯优化算法优化模型，从而对房价作出更加准确的评估。

本文的主要工作如下：

1)构建房价特征集，针对序列前向选择法的不足导致的特征冗余问题，提出一种融入Pearson相关系数的序列前向选择法，可以有效避免冗余特征，筛选出最优特征子集。

2)提出一种集成LightGBM模型，通过Bagging集成策略增强模型的泛化能力，并针对Bagging集成中采样比例的划分和个体学习器数量的选取等组合计算问题，利用贝叶斯优化算法得出最优解，从而提升模型性能。

1　特征选择与模型构建

1.1　特征选择

多源数据中往往包含了高维度的特征，不能直接用于模型的训练，因此需要对原始数据集进行特征选择。特征选择不仅可以防止模型过拟合，降低模型的泛化误差；还可以减少训练时间，降低模型开发成本，减少硬件资源损耗。

本文基于Pearson相关系数与序列前向选择法选择最优特征子集。首先利用Pearson相关系数过滤掉相关性较大的冗余特征，其计算过程如下：

$$r(\theta_i,\eta_i)=\frac{\sum_{i=1}^{n}(\theta_i-\overline{\theta})(\eta_i-\overline{\eta})}{\sqrt{\sum_{i=1}^{n}(\theta_i-\overline{\theta})^2}\sqrt{\sum_{i=1}^{n}(\eta_i-\overline{\eta})^2}} \tag{1}$$

其中：$r(\theta_i,\eta_i)$表示特征θ_i与特征η_i之间的相关系数；$\overline{\theta}$、$\overline{\eta}$分别表示特征θ_i和特征η_i的均值。相关系数的绝对值$\left|r(\theta_i,\eta_i)\right|$越大，两者的线性关联程度越强，当$\left|r(\theta_i,\eta_i)\right|>0.8$时，表示两个特征之间有极强的线性相关性[11]，则需要过滤掉冗余特征。

基于上述方法过滤掉冗余部分后，得到新的特征集合$Y=\{y_1,y_2,\cdots,y_n\}$，接下来的目标就是寻找最优的特征子集Y^*。本文利用序列前向选择法选择特征子集。其算法具体描述为，特征子集Y_k从空集开始，每次从特征集合中选择一个特征y加入特征子集Y_k，最终使得特征函数$J(Y^*)$达到最优。序列前向选择法流程描述如算法1所示。

算法1　序列前向选择法。

输入：特征集合$Y=\{y_1,y_2,\cdots,y_n\}$；

输出：最优特征子集Y^*。

1) 初始化特征子集$Y_0=\varnothing$和迭代次数$k=0$。

2) 每次迭代加入一个新特征$Y_{k+1}=Y_k+y$，并计算特征函数$\max\limits_{y\notin Y_k}J(Y_k+y)$。

3) 当特征函数满足$\max\limits_{y\notin Y_k}J(Y_k+y)>J(Y_k)$时，转至步骤2)；否则转至步骤4)。

4) 输出最优特征子集Y^*，迭代结束。

1.2　模型构建

集成学习(Ensemble learning)可以有效地提高模型的泛化能力，因此逐渐成为机器学习的研究热点，并被称为当前机器学习的研究方向之首[12]。

集成学习一般采用某种结合策略，构建并融合多个基学习器来完成学习任务[13]。按照基学习器的类型异同，集成学习通常分为同质集成和异质集成两大类[14]。在此基础上衍生出了各种集成方法。

Breiman[15]提出了一种Bagging的集成方法。该方法的主要思想是通过自助法(Bootstrap)从训练集中抽取N个训练子集，然后对这N个训练子集进行训练可以生成N个基学习器，最终结果由这N个基学习器投票或平均的方式得出，这样不仅提高了模型学习的精度，而且还可以降低过拟合的风险。

在Bagging的框架下，以决策树作为基学习器的随机森林[16]应运而生。由于随机森林在学习任务中展现出的良好性能，且能够容忍一定的异常数据和噪声，在信息技术、生物医学、经济管理学等诸多领域有着广泛的应用[17]。

本文借鉴随机森林的思想，提出一种基于贝叶斯优化的集成LightGBM模型。首先通过Bagging方法集成多个LightGBM，再结合贝叶斯优化算法优化模型，最后通过加权平均的方式获得最终输出。其实现方式如图1所示。

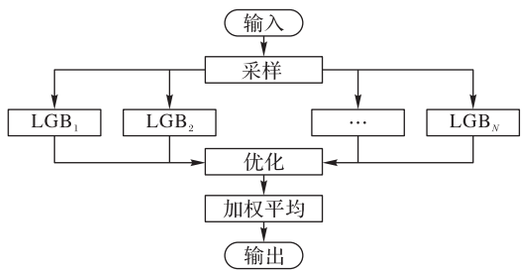

图1　集成LightGBM模型

Fig. 1　Ensemble LightGBM model

针对多个基学习器的集成问题，本文采用softmax函数为基学习器赋权，通过加权平均获得最终结果。

softmax函数通过归一化的方式，使每一个元素的取值都在(0,1)区间，且元素和为1，它能够凸显其中较大的元素，即为更准确的学习器赋予更大的权值。设第i个基学习器的输出结果的百分误差在10%以内的比例占g_i，则n个基学习器获得的权值可分别表示为S_i：

$$S_i=e^{g_i}\Big/\sum_{i=1}^{n}e^{g_i} \tag{2}$$

2 LightGBM 原理

LightGBM 是微软提出的一款开源的基于决策树的梯度提升框架，作为 Gradient Boosting 的改进版本，具有准确率高、训练效率高、支持并行和 GPU、使用内存小以及可以处理大规模数据[18]等优点。

2.1 Gradient Boosting

根据基学习器生成方式的不同，集成学习可以分为并行学习和串行学习。作为串行学习中最典型的代表，Boosting 算法又可分为 Adaboost 和 Gradient Boosting，它们的主要区别在于前者通过增加错分数据点的权重来提升模型，而后者通过计算负梯度来提升模型。

Gradient Boosting 的核心思想是利用损失函数的负梯度在当前模型 $f(x)=f_{j-1}(x)$ 的值近似替代残差。设训练样本为 $i(i=1,2,\cdots,n)$，迭代次数为 $j(j=1,2,\cdots,m)$，损失函数为 $L(y_i,f(x_i))$，则负梯度 r_{ij} 的计算公式如下：

$$r_{ij}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{j-1}(x)} \tag{3}$$

使用基学习器 $h_j(x)$ 拟合损失函数的负梯度 r，求出使损失函数最小的最佳拟合值 r_j：

$$r_j=\arg\min_r L(y_i,f_{j-1}(x_i)+rh_j(x_i)) \tag{4}$$

模型更新：

$$f_j(x)=f_{j-1}(x)+r_jh_j(x) \tag{5}$$

Gradient Boosting 在每轮迭代都会生成一个基学习器，通过多轮迭代，最终的强学习器 $F(x)$ 是由每轮产生的基学习器通过线性相加的方式求得：

$$F(x)=f_m(x) \tag{6}$$

2.2 改进策略

作为一款改进的轻量级 Gradient Boosting 算法，LightGBM 的核心思想是：直方图算法、带深度限制的叶子生长策略、直接支持类别特征、直方图特征优化、多线程优化、Cache 命中率优化。前两个特点有效地控制了模型的复杂度，实现了算法的轻量化，因此是本文尤其关注的。

直方图算法是通过把连续的浮点型特征离散化成 L 个整数，以构造一个宽度为 L 的直方图。遍历数据时，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，直方图累积了需要的统计量，然后从直方图的离散值中，寻找最优的分裂点。

传统的叶子生长策略对于同一层的叶子可以同时进行分裂，实际上很多叶子的分裂增益较低，没有必要分裂，这样带来了很多不必要的开销。对此 LightGBM 使用一种更加高效的叶子生长策略：每次从当前所有叶子中寻找分裂增益最大的一个叶子进行分裂，并设置一个最大深度限制。在保证高效的同时又防止了模型过拟合。

3 贝叶斯优化

贝叶斯优化算法是一种高效的优化算法，已经证明在一系列具有挑战性的优化问题上优于其他先进的优化算法。在数学上，可以统一将此问题描述为求解未知目标函数的全局最优解[19]：

$$x^*=\arg\max_{x\in X} f(x) \tag{7}$$

其中：x 表示待优化的参数；X 表示待优化的参数集合；$f(x)$ 表示目标函数。

在执行贝叶斯优化算法时有两个关键步骤。首先，必须选择一个先验函数来表示被优化函数的分布假设。为此，选择高斯过程，因为它具有灵活性和易处理性；其次，必须构建一个采集函数，用于从模型后验分布中确定下一个需要评估的点。

3.1 高斯过程

高斯过程是多维高斯分布在无限维随机过程上的扩展。它是通过均值函数和协方差函数定义的。

$$m(x)=E[f(x)] \tag{8}$$

$$k(x,x')=E\left[\left(f(x)-m(x)\right)\left(f(x')-m(x')\right)\right] \tag{9}$$

则高斯过程可以写作：

$$f(x)\sim GP\left(m(x),k(x,x')\right) \tag{10}$$

为了方便起见，通常将均值函数设为零。已知 $\left\{(x_i,f_i)\,|\,i=1,2,\cdots,n\right\}$，则存在一个高斯分布满足：

$$f\sim \mathrm{N}\left(0,\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})\right) \tag{11}$$

协方差矩阵 $\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})$ 和协方差函数 $k(x,x')$ 可以表示为：

$$\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})=\begin{pmatrix} k(x_1,x_1) & k(x_1,x_2) & \cdots & k(x_1,x_n)\\ k(x_2,x_1) & k(x_2,x_2) & \cdots & k(x_2,x_n)\\ \vdots & \vdots & & \vdots\\ k(x_n,x_1) & k(x_n,x_2) & \cdots & k(x_n,x_n)\end{pmatrix} \tag{12}$$

$$k(x,x')=\exp\left(-\frac{1}{2}|x-x'|^2\right) \tag{13}$$

协方差函数表示了函数的分布。因此，加入新样本 $\boldsymbol{X}^*$，并利用协方差矩阵生成一个新的高斯分布。

$$f^*\sim \mathrm{N}(0,\boldsymbol{K}(\boldsymbol{X}^*,\boldsymbol{X}^*)) \tag{14}$$

由高斯过程的性质可得，训练输出 f 和测试输出 f^* 的联合分布为：

$$\begin{pmatrix} f\\ f^*\end{pmatrix}\sim \mathrm{N}\left(0,\begin{bmatrix}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}^*)\\ \boldsymbol{K}(\boldsymbol{X}^*,\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}^*,\boldsymbol{X}^*)\end{bmatrix}\right) \tag{15}$$

则 f^* 的联合后验分布满足：

$$f^*|\boldsymbol{X}^*,\boldsymbol{X},f\sim \mathrm{N}(\boldsymbol{K}(\boldsymbol{X}^*,\boldsymbol{X})\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})^{-1}f,\\ \boldsymbol{K}(\boldsymbol{X}^*,\boldsymbol{X}^*)-\boldsymbol{K}(\boldsymbol{X}^*,\boldsymbol{X})\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}^*)) \tag{16}$$

通过评估均值和协方差矩阵，可以从联合后验分布中对函数值 f^* 进行采样[20]。

3.2 采样函数

通过采样函数确定下一个需要评估的点，可以减少迭代次数，降低评估成本。通常，对于采样点的选择从利用（exploitation）和探索（exploration）两个方面考虑。exploitation 就是根据当前的最优解，在其周围搜索，从而找到全局最优解；exploration 就是尽力探索未评估过的样本点，避免陷入局部最优解。

常用的采样函数有：EI（Expected Improvement）函数、POI（Probability of Improvement）函数以及 UCB（Upper Confidence Bound）函数。本次实验选取 UCB 函数作为采样函数，其数学表达式为：

$$UCB(x)=\mu(x)+\varepsilon\delta(x) \tag{17}$$

其中 $\mu(x)$ 和 $\delta(x)$ 分别是采用高斯过程所得到的目标函数联

合后验分布的均值和协方差。从式(17)可以看出，通过调节参数ε的大小，可以用来平衡采样点的选择[21]。

4　实验与结果分析

4.1　实验数据集

为验证模型的普适性和在真实场景中的准确性，本次实验分别使用了公开房价数据集和真实房价数据集。

公开房价数据集选取美国加州房价数据集，数据包含10个字段，其字段含义和数据类型如表1所示。

表1　加州房价数据集

Tab. 1　California house price dataset

字段含义	数据类型
经度	数值型
纬度	数值型
平均房龄	数值型
房间总数	数值型
卧室总数	数值型
人口总数	数值型
住户总数	数值型
收入中位数	数值型
房价中位数	数值型
离海距离	类别型

真实房价数据集来源于房产交易数据、地图软件应用程序接口(Application Program Interface，API)数据、公共交通数据以及城市地理信息数据，字段包括建筑面积、建筑年代、所处楼层、总楼层、户型、装修、物业费、停车位、容积率、绿化率、梯户比、经纬度、交通便捷程度、到中央商务区(Central Business District，CBD)距离、生活设施配套和教育配套，如表2所示。

表2　真实房价数据集

Tab. 2　Real house price dataset

具体字段	数据类型
建筑面积	数值型
建筑年代	数值型
所处楼层	类别型
总楼层	数值型
户型	类别型
装修	类别型
物业费	数值型
停车位	数值型
容积率	数值型
绿化率	数值型
梯户比	类别型
经纬度	数值型
交通便捷程度	类别型
到CBD距离	数值型
生活设施配套	类别型
教育配套	类别型

4.2　数据处理

在海量的原始数据中，存在着大批有缺失、有异常的数据，严重地影响到对数据潜在价值的挖掘。

一方面需要填充缺失值，过滤异常值。例如对于建筑面积、建筑年代、所处楼层等数据的部分缺失，用插值法进行填充；删除不合常理的极大或极小的异常数据等。另一方面是要使数据更平滑，从而让数据更好地适应模型。例如对房价数据乘以对数函数，使得数据近似服从正态分布。

此外，由于实验数据具有多维度，因此需要对数据进行规范化，目的是消除不同数据之间取值范围和量纲的影响，其公式如下所示：

$$d^{*} = (d - \mu)/\delta \tag{18}$$

其中：d表示特征数据；μ表示数据的均值；δ表示数据的方差。

4.3　评价指标

为了合理评价模型的综合性能，本文分别构建对数平均绝对误差(Mean Absolute Logarithmic Error，MALE)和对数均方根误差(Root Mean Square Logarithmic Error，RMSLE)作为模型的综合评价指标。MALE能更好地反映观测值误差的实际情况，RMSLE用来衡量观测值和真实值之间的偏差，两者的研究目的不同，但是计算过程相似，公式定义为：

$$MALE = \frac{1}{n}\sum_{i=1}^{n}\left|\ln\left(p_i + 1\right) - \ln\left(\widehat{p_i} + 1\right)\right| \tag{19}$$

$$RMSLE = \sqrt{\frac{\sum_{i=1}^{n}\left[\ln\left(p_i + 1\right) - \ln\left(\widehat{p_i} + 1\right)\right]^2}{n}} \tag{20}$$

其中：p_i表示实际的房价；$\widehat{p_i}$表示模型输出的房价。

4.4　模型对比

将集成LightGBM模型与当前公开研究中提及的经典模型进行对比实验，下面对各个模型进行简要介绍。

线性回归　利用线性预测函数，对自变量和因变量进行建模的一种回归分析。当只有一个自变量时称为一元线性回归，当自变量大于一个时称为多元线性回归。

多项式回归　利用多项式的回归分析方法，对自变量和因变量进行建模，通过增加自变量的高次项对因变量进行拟合，能够解决一些非线性问题。

K近邻　计算该样本与所有训练样本的距离，然后找出与它最接近的k个样本，将样本分到离它最接近的样本所属的类中。

BP神经网络　利用误差反向传播算法训练的多层前馈神经网络，是目前应用最广泛的神经网络模型之一。

支持向量机　通过寻找一个超平面来对样本进行分割，它不仅能正确地对每一个样本进行分类，并要使每一类样本中离超平面最近的样本与超平面之间的距离尽可能远。

随机森林　利用随机有放回采样得到的样本训练多棵决策树，决策树的每个节点在训练时只用了样本无放回抽样的部分特征，最后用这些决策树的预测结果进行投票或平均。

本次实验选取加州房价数据集和真实房价数据集作为训练集，对各类模型进行训练，结果如表3所示。

表3　各类模型对比结果

Tab. 3　Comparison results of various models

模型	加州房价数据集		真实房价数据集	
	MALE	RMSLE	MALE	RMSLE
线性回归	0.247	0.326	0.157	0.206
多项式回归	0.187	0.259	0.115	0.184
K近邻	0.174	0.236	0.091	0.136
BP神经网络	0.162	0.227	0.078	0.113
支持向量机	0.202	0.279	0.112	0.156
随机森林	0.164	0.230	0.071	0.103
LightGBM	0.151	0.218	0.069	0.098
集成LightGBM	0.143	0.207	0.065	0.092

不难看出，本文提出的集成LightGBM模型性能明显优于*K*NN、SVM这类单一模型，随机森林、LightGBM这类集成模型，以及BP神经网络这类深度学习模型，进一步验证了集成学习在机器学习中展现的优越性。

4.5 参数敏感性测试

由于基学习器的个数和采样比例决定着集成效果的好坏，因此在加州房价数据集上对模型的参数组合问题作敏感性测试，参数取值如表4所示。

表4 参数敏感性测试取值

Tab. 4 Values of parameter sensitivity test

模型参数	参数取值
基学习器个数	(10,20,30,40,50,60,70,80,90,100)
采样比例	(0.5,0.6,0.7,0.8,0.9)

本次实验选取均方误差作为评价标准，均方误差越小，算法准确率越高。由表4可得，共有50种参数组合。若随着参数组合变动，均方误差一直处于上下波动状态，则认为模型对参数敏感；若均方误差在某个参数组合之后趋于平稳，则认为模型对参数不敏感。测试结果如图2所示。

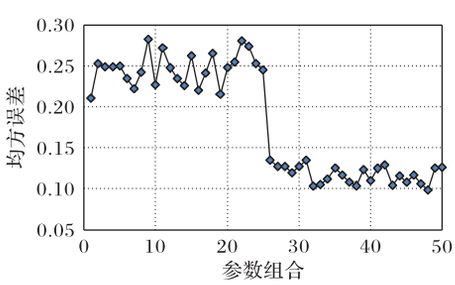

图2 参数敏感性测试

Fig. 2 Parameter sensitivity test

经过50种参数组合模型依然无法达到最优，由此证实，参数的优劣极大地影响了模型性能。

4.6 参数优化

在真实房价数据集的基础上，分别使用网格搜索法和贝叶斯优化算法，对采样比例和基学习器数量进行优化。设采样比例在0.5~0.9，基学习器数量在10~100，结果如表5所示。

表5 模型优化结果

Tab. 5 Model optimization results

优化方法	迭代次数	MALE	RMSLE
网格搜索	150	0.061	0.087
贝叶斯优化	100	0.058	0.083

显然，贝叶斯优化在更少的迭代次数中获得更优的结果，在参数组合寻优问题上优于传统的网格搜索，能够在实际的应用中减少时间开销，提升模型性能。

4.7 预测结果

基于本文提出的模型在真实场景下对房价进行智能评估，将真实房价数据集按照9∶1的比例随机分为训练集和测试集，其输出结果与真实房价的拟合曲线如图3所示。

从图3可以看出，上述构建的集成学习模型输出的房价与实际的房价能够较为准确地拟合。

与此同时，为了更加真实地反映输出结果的可信度，本文通过百分误差来衡量输出值与真实值之间的偏差，其计算过程如下：

$$E = \left| \frac{p_i - \widehat{p_i}}{p_i} \right| \cdot 100\% \tag{21}$$

其中：E表示百分误差；p_i表示实际的房价；$\widehat{p_i}$表示输出的房价。

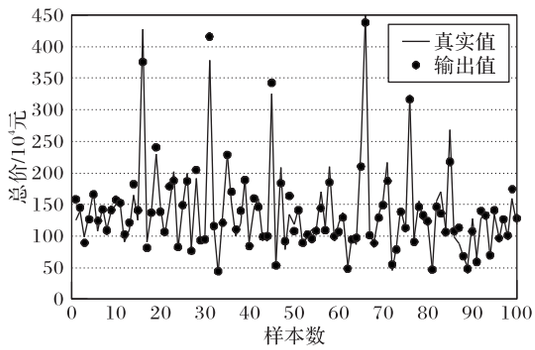

图3 模型输出结果与真实值的拟合曲线

Fig. 3 Fitted curve between model output and real value

将本文构建的集成模型与子模型进行可信度分析。定义如下：输出结果与实际房价的百分误差在10%以内，具有较高的可信度；输出结果与实际房价的百分误差在10%~20%，可信度中等；输出结果与实际房价的百分误差在20%以上，可信度较低。分析结果如图4所示。

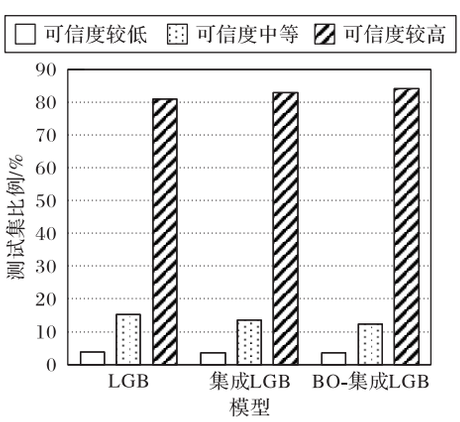

图4 可信度分析

Fig. 4 Reliability analysis

对于可信度较高的输出结果，在实际的应用中能够准确地反映房价，具有很高的应用价值；可信度中等的输出结果，在一定程度上也能够作为房价的参考。由图4可得，本文提出的基于贝叶斯优化的集成LightGBM模型(BO-集成LGB)，较LightGBM模型(LGB)在精度上提升3.15个百分点，其96.46%的输出结果都能在真实场景中发挥它的价值，进一步体现了该模型在评估准确性上的优势。

综合上述分析证明，采用集成学习和贝叶斯优化算法对LightGBM的改进是有效的，能够较为准确地评估房价，在实际的房价评估中具有一定的指导意义。

5 结语

随着信息技术的飞速发展，大数据、人工智能为企业、社会、甚至是国家带来了前所未有的机遇。本文在多源数据的基础上，提出了一种基于贝叶斯优化的集成LightGBM模型。实验表明，所提模型准确率优于*K*NN、SVM这类单一模型，随机森林、LightGBM这类集成模型，以及BP神经网络这类深度

学习模型,房价评估结果也与实际值比较接近,进而体现了数据挖掘的意义,实现了海量数据的价值。

房地产市场的特殊性,时间、人文、经济环境等因素也会不同程度地影响房价。在未来的工作中,将结合我国的基本国情,对影响房价的指标进一步细化,充分提取潜在的影响因子,使评估结果更加准确。

参考文献 (References)

[1] LIU G, ZOHG X. Research of second-hand real estate price forecasting based on data mining[C]// Proceedings of the IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference. Piscataway: IEEE, 2017: 1675-1679.

[2] PHAN T D. Housing price prediction using machine learning algorithms: the case of Melbourne city, Australia[C]// Proceedings of the 2018 International Conference on Machine Learning and Data Engineering. Piscataway: IEEE, 2018: 35-42.

[3] FENG Y, JONES K. Comparing multilevel modelling and artificial neural networks in house price prediction[C]// Proceedings of the 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services. Piscataway: IEEE, 2015: 108-114.

[4] MUKHLISHIN M F, SAPUTRA R, WIBOWO A. Predicting house sale price using fuzzy logic, artificial neural network and k-nearest neighbor[C]// Proceedings of the 1st International Conference on Informatics and Computational Sciences. Piscataway: IEEE, 2017: 171-176.

[5] LU S, LI Z, QIN ZHEN, et al. A hybrid regression technique for house price prediction [C]// Proceedings of the 2017 IEEE International Conference on Industrial Engineering and Engineering Management. Piscataway: IEEE, 2017: 319-323.

[6] 王昕睿. 基于机器学习的房产智能自动评估模型的研究与系统实现[D]. 北京:北京邮电大学, 2019: 37.(WANG X R. Research and system implementation of intelligent automatic evaluation model for real estate based on machine learning[D]. Beijing: Beijing University of Posts and Telecommunications, 2019: 37.)

[7] 刘燕云. 基于兰州市二手房价评估模型研究[D]. 兰州:兰州大学, 2019:37.(LIU Y Y. Research on Lanzhou second-hand house price evaluation model [D]. Lanzhou: Lanzhou University, 2019: 37.)

[8] 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.(CHEN M, LI Y B. Automatic evaluation of second-hand house prices in Wuhan based on hedonic price theory and neural network[J]. Urban Geotechnical Investigation and Surveying, 2018(4): 21-24.)

[9] 王海泉. 武汉市二手房价格评估研究[D]. 武汉:华中师范大学, 2018:35-49.(WANG H Q. Study on the price evaluation of second-hand houses in Wuhan[D]. Wuhan: Central China Normal University, 2018: 35-49.)

[10] 李恒凯,柯江晨,王秀丽. 融GIS和BP神经网络的住宅房产评估模型[J]. 测绘科学, 2018, 43(8):104-109.(LI H K, KE J C, WANG X L. Evaluation model of residential property based on GIS and BP neural network model method [J]. Science of Surveying and Mapping, 2018, 43(8): 104-109.)

[11] 张良均,王路,谭立云,等. Python数据分析与挖掘实战[M]. 北京:机械工业出版社, 2016:48(ZHANG L J, WANG L, TAN L Y, et al. Python Practice of Data Analysis and Mining[M]. Beijing: China Machine Press, 2016: 48.)

[12] DIETTERICH T G. Machine learning research: four current directions[J]. AI Magazine, 1997, 18(4): 97-136.

[13] 周志华. 机器学习[M]. 北京:清华大学出版社, 2016:171-173.(ZHOU Z H. Machine Learning[M]. Beijing: Tsinghua University Press, 2016: 171-173.)

[14] 周钢,郭福亮. 集成学习方法研究[J]. 计算机技术与自动化, 2018, 37(4): 148-153.(ZHOU G, GUO F L. Research on ensemble learning[J]. Computing Technology and Automation, 2018, 37(4): 148-153.)

[15] BREIMAN L. Bagging predicators[J]. Machine Learning, 1996, 24(2): 123-140.

[16] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[17] 方匡南,吴久彬,宋建平,等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3):32-38.(FANG K N, WU J B, SONG J P, et al. A review of random forests [J]. Statistics and Information Forum, 2011, 26(3): 32-38.)

[18] GUO L, QI M, FINLEY T, et al. LightGBM: a highly efficient gradient boosting decision tree [C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 3146-3154.

[19] SHAHRIARI B, SWERSKY K, WANG Z, et al. Taking the human out of the loop: a review of Bayesian optimization [J]. Proceedings of the IEEE, 2015, 104(1): 148-175.

[20] RASMUSSEN C E, WILLIAMS C K I. Gaussian Processes for Machine Learning[M]. Cambridge: MIT Press, 2005: 13-16.

[21] SNOEK J, LAROCHELLE H, ADAMS R P. Practical Bayesian optimization of machine learning algorithms[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2012: 2951-2959.

This work is partially supported by the Ministry of Education-China Mobile Scientific Research Fund (MCM20170203), the Chongqing Natural Science Foundation (cstc2018jcyjAX0587), the Chongqing Technology Innovation and Application Demonstration (Industry Key Research and Development) Project (cstc2018jszx-cyzdX0124).

GU Tong, born in 1995, M. S. candidate. His research interests include machine learning, data mining.

XU Guoliang, born in 1973, Ph. D., professor. His research interests include photoelectric sensing and detection, communication network design and planning, big data analysis and mining.

LI Wanlin, born in 1963, Ph. D., professor. His research interests include next-generation network technology, autonomous driving, Internet of vehicles, mobile big data.

LI Jiahao, born in 1994, M. S. candidate. His research interests include data mining.

WANG Zhiyuan, born in 1995, M. S. candidate. His research interests include data mining.

LUO Jiangtao, born in 1971, Ph. D., professor. His research interests include mobile big data, next-generation network technology, communication network testing and optimization.