GAN-BPM:基于 GAN 的子市场划分房屋定价模型

唐朝君, 章 平,包象琳,徐晓峰
安徽工程大学 计算机与信息学院,安徽 芜湖 241000

摘 要:目的 针对城市房屋市场数据中存在着分布和类别不平衡以及特征价格模型(Hedonic Price Model,HPM)面对房价与特征关系可解释性不足的问题,提出一种基于对抗生成网络(Generative Adversarial Net,GAN)的子市场划分房屋定价模型。**方法** 在改进子市场划分房屋定价模型时,首先引入 GAN 作为数据增强技术模块,生成具有多样性和逼真度的合成样本,增加样本的多样性,提高模型的泛化能力;接着将 GAN 数据增强和子市场划分的房屋定价模型相结合;最后,依据房屋所属子市场的概率预测房价,并分析对房价的关键影响因素,提升预测精度和可解释性。**结果** 将模型与 5 个现有模型以及未加入 GAN 的子市场划分房屋定价模型,从平均绝对百分比误差、平均绝对误差和均方根误差 3 个方面进行对比;通过使用杭州市 2020 年的房产数据,对模型的算法性能进行测试。**结论** 实验证明引入 GAN 数据增强技术模块后,模型在房地产价格预测方面优于其他对比模型,并且具有可解释性的优点。

关键词:房价评估;GAN;数据增强;房产数据;位置特征;子市场
中图分类号:F299.23;F224 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0006.014

GAN-BPM: Sub-market Division Housing Pricing Model Based on GAN

TANG Chaojun, ZHANG Ping, BAO Xianglin, XU Xiaofeng
School of Computer and Information, Anhui Polytechnic University, Wuhu 241000, Anhui, China

Abstract: Objective Addressing issues of distribution and class imbalance in urban housing market data, along with insufficient interpretability of the Hedonic Price Model (HPM) in explaining the relationship between house price and features, a housing pricing model for sub-market division based on generative adversarial networks (GAN) is proposed. **Methods** In improving the sub-market division housing pricing model, GAN is first introduced as a technical module for data enhancement to generate synthetic samples with diversity and fidelity to increase the diversity of the samples and improve the generalization ability of the model. Subsequently, GAN-based data augmentation is integrated with the sub-market division to develop a housing pricing model. Finally, housing prices are predicted based on the probability of the submarkets to which the houses belong and key influencing factors on the house prices are analyzed to improve the prediction accuracy and interpretability. **Results** The proposed model is compared with five existing models and a sub-market division housing pricing model without GAN based on average absolute percentage error, average absolute error,

收稿日期:2023-11-07 **修回日期:**2024-01-15 **文章编号:**1672-058X(2025)06-0105-10
基金项目:安徽省自然科学基金(2108085QF268,2108085QF264);安徽工程大学校级科研项目(XJKY2022154).
作者简介:唐朝君(1998—),男,安徽六安人,硕士研究生,从事房产评估研究.
通信作者:章平(1982—),男,安徽芜湖人,博士,副教授,硕士生导师,从事无线定位、城市计算、物联网研究. Email:pingzhang@ahpu.edu.cn.
引用格式:唐朝君, 章平,包象琳,等. GAN-BPM:基于 GAN 的子市场划分房屋定价模型[J]. 重庆工商大学学报(自然科学版),2025,42(6):105-114.
TANG Chaojun, ZHANG Ping, BAO Xianglin, et al. GAN-BPM: Sub-market division housing pricing model based on GAN[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(6): 105-114.

and root mean square error. Algorithm performance is tested using 2020 real estate data from Hangzhou. **Conclusion** Experimental results demonstrate that after introducing the data augmentation technology module based on GAN, this model is superior to comparative models in predicting real estate prices and has the advantage of interpretability.

Keywords: housing price evaluation; GAN; data augmentation; real estate data; location features; sub-market

1 引 言

房地产市场作为全球经济中的重要组成部分,一直以来都受到广泛的关注。房屋定价是房地产市场中的核心问题之一,准确的房屋定价对于市场参与者、政府监管机构和金融机构具有重要意义。然而,由于房地产市场的复杂性和不确定性,房屋定价往往面临着挑战和局限性。

随着城市化进程的不断加速,对于城市房屋市场数据的需求与日俱增。城市房屋市场数据在城市规划、房地产市场研究、投资决策和房屋价格预测等领域具有重要的应用价值。Sandeep 等[1]的研究为住房购买者提供房屋价格信息和房屋市场趋势,协助购房者进行购房选择。房地产价值受多种因素影响,如地理位置,不同位置的房地产,价格差异较大[2];核心商圈、风景区、高铁站等对房地产价格亦有不同程度的影响[3]。

为了克服数据获取和多样性不足的困难,近年来,生成对抗网络(Generative Adversarial Net,GAN)已经成为一种强大的工具,被广泛应用于各种数据生成任务。GAN 作为一种强大的生成模型,可以做到利用高质量图像生成进行数据增强[4]。GAN 通过训练生成器和判别器相互博弈的方式,生成逼真的合成样本,从而扩充了原始数据集的规模和多样性。现有的以 GAN 作为数据增强的方法中,较少有研究将 GAN 作为数据增强技术,来对房地产数据进行扩充。通过引入 GAN 数据增强技术,房地产市场的分析和聚类任务可以受益于更多的样本和更好的数据分布,从而提高模型的性能和鲁棒性。

另一方面,房屋市场数据通常包含多个子市场,每个子市场具有不同的特征和行为模式。传统聚类方法往往忽略了这种潜在的子市场结构,导致对整个市场的分析结果缺乏准确性和解释性。为了更好地捕捉市场的内在结构和特点,子市场划分的贝叶斯概率模型能够将市场数据分割为多个子市场,并对每个子市场的分布进行建模,以此来构建房价预测模型。但现有房价预测模型研究较少用到数据增强技术。

本文的贡献在于使用基于 GAN 的数据增强技术模块,将其和子市场划分的贝叶斯概率模型结合作为房屋定价模型。本文将 GAN 用于生成合成样本,以扩充原始数据集的规模和多样性,并通过子市场划分的贝叶斯概率模型对市场数据进行分割,从而捕捉市场的内在结构和特点。在真实的市场数据集上进行广泛的实验评估,并与传统方法进行比较。实验结果表明:本文方法能够提高模型的准确度和鲁棒性,同时提供更可解释的结果,为商业决策提供更有价值的信息。

2 相关工作

现代社会,房地产已经成为人民生活和国家经济最重要的话题之一。合理评估房屋特征对房地产价格的影响,对城市规划和社会生活具有重要意义。为了提高房价预测精度,很多学者都进行了深入研究。Sun 等[5]在评估中使用遗传算法来优化神经网络模型,通过遗传神经网络选择最合适的权重和阈值,有效地降低了评估误差;Xu 等[6]利用信息增益比的方法对特征进行加权、排序,在有限混合模型的基础上建立了回归模型。

Wu 等[7]的研究聚焦于房地产市场中空间连续性对住房价格分类的影响,提出一种基于主成分分析和聚类分析的数据驱动模型。他们利用房屋地理位置的信息,将整个空间划分为不同的子市场。这种划分使他们能够更好地理解和预测每个子市场中的房价。在此基础上,Liu 等[8]进一步提出一种基于建筑环境和房屋基础特征的子市场建模方法,以提高房价预测的精度;秦心静等[9]认为建筑环境和房屋基础特征对房价具有重要影响,因此采用这些特征对每个子市场进行建模。通过利用子市场效应,他们能够提升房价预测的准确性,并通过概率层次聚类的方法推断市场的层次结构。这些研究为房价预测提供了一种综合的数据驱动方法,通过考虑空间连续性和特征影响,能够更准确地刻画房地产市场的子市场,并提高预测模型的精度。

数据增强是一种常用的技术,用于扩充训练数据

集的大小和多样性,以改善机器学习模型的性能和泛化能力。通过对原始数据应用一系列变换和扰动,可以生成新的样本,这些样本在保持真实性的同时增加了数据的多样性,也是许多科研人员一直在研究的方向。

Wu 等[10]发现基因表达谱数据对于癌症诊断等各种任务的效率较低,因为其样本量小且维度高。条件谱归一化生成对抗性网络(CSN-GAN)可以生成指定标签的高质量样本,直接应用 CSN-GAN 生成基因表达谱数据将面临数据分布不正确和严重过拟合的问题。因此,他们提出了一种基于条件 SN-GAN 的基因表达谱增强方法,该方法具有可学习的输出层函数。通过将生成模型的强边界限制改变为弱限制,生成的数据分布尽可能接近真实样本分布。

异物悬浮是一种小概率事件,现有样本较少。使用 CNN 进行目标分类检测存在样本不足或样本不平衡的问题。针对 CNN 工程应用中经常出现的上述问题,Dou 等[11]提出一种基于 GAN 的数据增强算法。使用 GAN 生成的特征图和原始数据来训练预训练模型的分类层,从而达到数据增强和平衡样本的目的,增强模型的分类能力。

3　基于 GAN 的数据增强技术

生成器从房屋数据中捕获内部分布模式,并生成一组具有相同分布的房屋数据。选择长短期记忆神经(Long Short-Term Memory,LSTM)网络作为鉴别器,以区分实际的房屋数据和生成器生成的数据。在这个任务中,本文将鉴别器视为一个由另一个神经网络组成的损失函数[12]。

3.1　GAN 单元

GAN 单元是传统卷积神经网络的一种生成性变体,用于基于图像的深度学习。通过对抗性训练,GAN 能够生成与训练数据类似的样本。

网络的生成性质使得 GAN 在具有多个维度、低相关性或稀疏性的金融数据中尤为有效。GAN 是一个无监督学习模型,用于预测网络的过程。它不需要真实的训练数据作为标准,而是通过生成器模块利用随机抽样的输入变量来进行预测。在金融预测中,生成模型在高维度、稀疏数据上提供了更高质量的预测效果。

GAN 框架中有两个网络:生成器和判别器。在训练过程中,生成器会逐渐从一些随机噪声中学习输入数据集的关键特征;判别器则提高了对输入数据集进行“真实”分类以及对生成网络生成的数据进行“伪造”分类的能力。

生成器和判别器在整个训练过程中进行一个双方博弈的极小极大游戏。判别器力求将损失最小化,以准确地对真实样本和生成样本进行分类。生成器则试图最大化判别器损失,如式(1)所示,其中 $D(X_{\text{real},i})$ 代表真实数据,$D(X_{\text{fake},i})$ 代表生成数据。

$$\min_G \max_D V(G,D)= E[\log D(X_{\text{real},i})]+E[\log(1-D(X_{\text{fake},i}))] \quad (1)$$

3.2　算　法

在 GAN 架构中,采用 LSTM 作为鉴别器,系统架构图如图 1 所示。

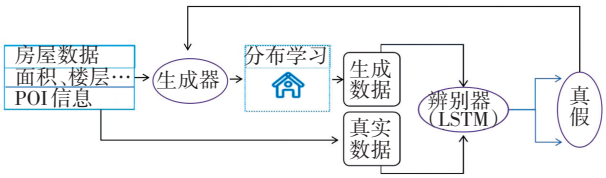


图 1　GAN 结构示意图

Fig. 1　GAN structure diagram

在上述体系结构中,有两个关键点。首先,生成器从随机噪声中逐步学习输入数据集的基本特征,根据真实房地产房屋数据,挖掘其数据分布,以生成相同分布的数据;其次,LSTM 设计的鉴别器,可以将输入数据集分类为“真实”,将生成器生成的数据分类为“伪造”。

LSTM 使用记忆来加强当前的决策。这 3 个控制门决定了存储空间。使用内存输出门决定了该部分是否被附加到输出上。它使用 sigmod 函数来指示是否添加它,如图 2 所示。

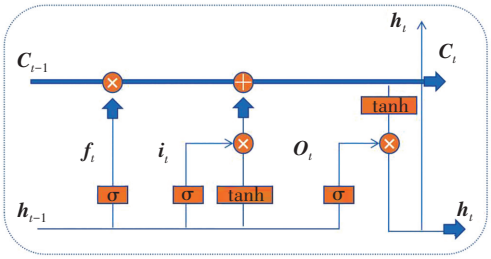


图 2　LSTM 结构示意图

Fig. 2　Schematic diagram of LSTM structure

图 2 显示了 LSTM 结构,其中每一行代表从一个节点到另一个节点的一个向量。橙色的圆圈执行运算,如向量的加法或乘法。这 4 层是 3 个 sigmod 层和一个 tanh 层,用橙色方框表示。一个存储单元的输出与后续单元的输入连接起来。在 LSTM 中,新的信息可以通过称为门的结构添加到单元状态($\boldsymbol{C}_t$)中,门是 LSTM 单元的重要部分。

忘记门决定忘记或保持单元状态的信息,如下所示。

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{2}$$

通过使用 sigmoid 函数,可以对先前隐藏状态和当前输入信息进行处理。sigmoid 函数的输出值范围在 0 到 1 之间,用于控制数据的保留或遗忘。具体而言,忘记门和输入门在这一过程中起着关键作用。

忘记门利用 sigmoid 函数处理先前隐藏状态和当前输入信息,输出一个值,用于决定需要遗忘多少旧的单元格状态信息。当输出值接近于 0 时,表示需要遗忘的信息较多;而当输出值接近于 1 时,表示需要保留的信息较多。

与此同时,输入门也使用 sigmoid 函数对先前隐藏状态和当前输入信息进行处理,输出一个值,用于确定哪些数据应该添加到单元格状态中。输出值接近于 0 表示对应的数据不重要,而接近于 1 则表示数据重要。

最终,根据忘记门和输入门的输出值,单元格状态进行更新。忘记门的输出值与旧的单元格状态进行乘法运算,以决定需要遗忘的信息。输入门的输出值与调整后的先前隐藏状态和当前输入信息进行乘法运算,然后将两部分相加,得到新的单元格状态。通过这样的方式,单元格状态可以动态地更新和调整,以适应不同的输入情况。

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i) \tag{3}$$

$$\widetilde{\boldsymbol{C}}_t=\tanh(\boldsymbol{W}_C[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_C) \tag{4}$$

新单元状态由式(3)组成,它表示新信息,根据本文决定更新状态的程度计算,如下所示。

$$\boldsymbol{C}_t=\boldsymbol{f}_t\cdot\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\cdot\widetilde{\boldsymbol{C}}_t \tag{5}$$

输出门的作用是确定下一个隐藏状态的值,并且隐藏状态包含之前输入的相关信息。

通常,输出门使用 tanh 函数来决定是否将长期内存添加到输出中。tanh 函数的输出范围是[-1, 1],因此它可以控制是否保留或删除长期内存中的信息。具体而言,输出门的输出值与当前单元格状态输入到 tanh 函数中,得到一个介于-1 和 1 之间的值。这个值与输出门的输出值相乘,得到下一个隐藏状态的值。

通过这样的方式,隐藏状态不仅包含了当前输入的信息,还包含了之前输入的相关信息。隐藏状态可以作为下一步的输出,也可以传递给下一层网络进行进一步处理。这样,模型可以利用隐藏状态中蕴含的历史信息来进行预测和决策。

综合这 3 个门,设计式如下。

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o) \tag{6}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\cdot\tanh(\boldsymbol{C}_t) \tag{7}$$

其中,$\boldsymbol{o}_t$ 是输出门,$\boldsymbol{h}_{t-1}$ 是 $t-1$ 时刻的隐藏状态,$\boldsymbol{W}_o$ 是要训练的权值,$\boldsymbol{x}_t$ 是 t 时刻的输入,$\boldsymbol{b}_o$ 是偏差值,$\boldsymbol{h}_t$ 是 t 时刻的隐藏状态,$\boldsymbol{C}_t$ 是 t 时刻的单位状态。

3.3 人工约束

在基于 GAN 的城市房屋市场数据增强技术中,人工约束部分是关键的组成部分之一。通过人工约束,可以确保生成的城市房屋数据符合实际的房屋设计,如房屋布局、房屋面积等。这些约束可以帮助生成的数据更加合理、可行,并具有与真实房屋数据相似的特征分布。并且通过加入人工约束,可以对数据集中部分特征较少或缺少的数据,进行补充,以丰富数据样本,增加数据集的多样性。

在人工约束中,一种常见的方法是使用地理信息系统(GIS)数据来提供城市规划的约束。GIS 数据包括道路网络、地块用途、建筑高度限制等信息,这些信息可以作为生成模型的输入或约束条件。通过将这些信息与 GAN 模型结合,可以生成符合实际城市布局规划的房屋数据。

另一种方法是引入专家知识或规则来约束生成模型的输出。城市规划专家可以提供他们的经验和见解,定义一些规则和约束条件,例如合理的房屋布局等。这些约束条件可以通过损失函数的设计或生成模型的网络结构进行整合,以确保生成的房屋数据满足这些规则和约束。

同时,还可以利用生成模型的条件输入来引入人工约束。例如,将特定的城市规划要求作为条件输入,例如要求生成具有特定功能的房屋(商业建筑、住宅等),或者要求在特定区域生成具有特定特征的房屋(如高密度区域、低密度区域等)。这些条件输入可以指导生成模型生成符合特定规划要求的房屋数据。

4 房屋定价模型

4.1 模型构建

基于房屋位置利用高德地图兴趣点(Point of Interest, POI)数据获取周边建筑类型、学校和交通设施等环境特征。并利用房屋位置与房屋特征构建基于子市场划分的贝叶斯概率房屋定价模型,模型结构如图 3 所示。

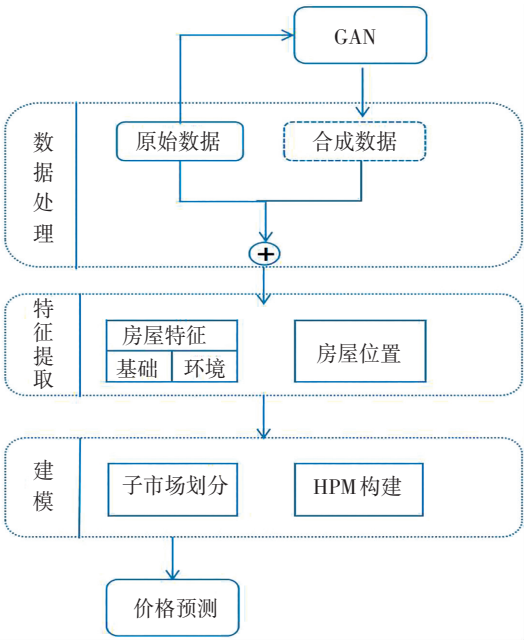


图 3　GAN-BPM 结构示意图

Fig. 3　GAN-BPM structure diagram

4.2　模型假设

本文所提出的房屋定价模型中,房屋特征、房屋位置以及房屋价格是可观测的,而房屋价格与房屋特征和位置的关系都依赖于子市场。

如图 4 所示,房屋的位置($\boldsymbol{P}$)决定了房屋所属的次级市场($\boldsymbol{S}$),而房屋本身的特征属性($\boldsymbol{R}$)与所属子市场决定了房屋本身的价格($\boldsymbol{Y}$)。房屋的特征主要包括基本特征(面积、楼层等)以及相关的 POI 信息。

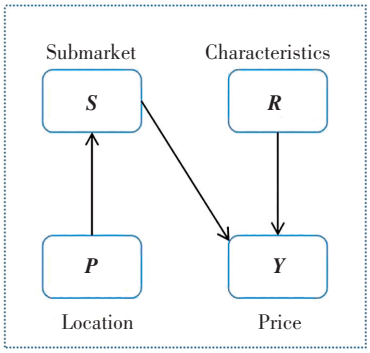


图 4　概率模型关系示意图

Fig. 4　Diagram of probability model relationship

本文将房屋所属的子市场视为随机变量,相关符号定义由表 1 所示。先验概率由参数为 a 的多点分布概率向量表示,反映了不同子市场房产数量的比例。房屋位置与子市场之间的关系使用高斯分布,参数 μ_{s_n} 和 $\sum_{s_n}$ 表示子市场的位置中心和范围大小。房屋价格由所属子市场和房屋特征共同确定,通过子市场内部的特征价格模型 $f_{s_n}(r_n)$ 来描述价格的浮动范围 $\sigma^2_{s_n}$。

表 1　符号定义

Table 1　Notations

符号	维数	描述
$\boldsymbol{R}$	35×N	房屋特征
$\boldsymbol{Y}$	N	房屋价格
$\boldsymbol{P}$	2×N	房屋位置
$\boldsymbol{S}$	N	所属子市场
K	1	子市场数量

对第 n 套住房,$n \in \{1,2,\cdots,N\}$,如式(8)所示:

$$\begin{aligned} s_n &\sim \text{Categorical}(a) \\ p_n &\sim N(p|\mu_{s_n},\Sigma_{s_n}) \\ y_n &\sim N(y|f_{s_n}(r_n),\sigma^2_{s_n}) \end{aligned} \tag{8}$$

该模型使用基于异质性相关标准的贝叶斯网络结构来划分房屋位置的子市场。子市场的位置和范围是根据空间邻近性和可替代性确定的。空间邻近性表示同一子市场内房屋位置的临近性,假设每个子市场都有一个中心点,高斯分布使得同一子市场的房屋位置围绕中心点分布。可替代性指的是相同子市场中房屋特征转售机制的相似性,使用 HPM 回归价格与观测价格之间的方差来衡量可替代性。

总而言之,房屋价格由房屋特征和所属子市场决定。每个子市场都有唯一的 HPM(本文使用的是梯度推进回归树)与之对应。相较于线性模型,该模型具有较低的偏差,并且通过部分依赖图的概念,具有良好的解释性。

梯度推进回归树(Gradient Boosting Decision Tree,GBDT)是一种集成学习算法,用于解决回归和分类问题。它通过迭代训练一系列决策树模型,并将它们组合成一个强大的预测模型。

在 GBDT 中,每棵决策树都是通过拟合前一棵树的残差(预测值与真实值之间的差异)来构建的。在每一轮迭代中,新的决策树被训练来预测之前模型的残差,然后将其加到模型中,逐步减少残差的误差。

为了避免过拟合,GBDT 使用了一种称为梯度下降的技术。在每一轮迭代中,通过计算损失函数对模型预测的梯度,来调整树的参数,使得下一棵树能够更好地拟合残差。这样,通过多轮迭代,每棵树都在前一棵树的基础上进一步优化预测结果,最终得到一个强大的集成模型。

GBDT 具有许多优点,包括对非线性关系的建模能力强,对异常值和噪声数据具有较好的鲁棒性,并且在特征工程方面相对较少的预处理需求。它在许多领域都取得了很好的应用效果,如预测问题、排名问题和异常检测等。

4.3 超参数估计

完成模型构造后,采用经验贝叶斯方法进行超参数估计,超参数有 $\theta=\{a_k,\mu_k,\Sigma_k,f_k,\sigma_k^2\}_{k=1}^K$、参数(潜在变量)$\boldsymbol{S}$、观测数据 $D=\{(p_n,y_n)\}$。给定观测数据和子市场数量 K 的情况下,使用最大似然原理估计 θ。D 的对数似然估计如式(9)所示:

$$L(\theta)=\log P_r(D\mid\theta,K)=\sum_{n=1}^{N}\log\left(\frac{\sum_{k=1}^{K}P_r(p_n,y_n\mid s_n=k,\theta,K)}{P_r(s_n=k\mid\theta,K)}\right) \tag{9}$$

为了最大化似然函数,使用期望最大化算法(EM 算法)评估超参数 θ,推导过程如式(10)和式(11)所示:

$$Q(\theta\mid\hat{\theta}^{(t)})=E_s[\log P_r(D,S\mid\theta)\mid D,\hat{\theta}^{(t)}]=\sum_s P_r(S\mid D,\hat{\theta}^{(t)})\log P_r(D,S\mid\theta) \tag{10}$$

$$\hat{\theta}^{(t+1)}=\underset{\theta}{\operatorname{argmax}}\,Q(\theta\mid\hat{\theta}^{(t)}) \tag{11}$$

EM 算法分为两步,期望(E 步)如式(12)所示,利用对超参数的现有估计值,评估所属子市场的后验概率。

$$\gamma(s_n)=P_r(k\mid p_n,y_n,\theta^{(t)},K)=\frac{P_r(k\mid\theta^{(t)},K)\,P_r(p_n,y_n\mid k,\theta^{(t)},K)}{\sum_{i=1}^{K}a_i^{(t)}\,P_r(y_n\mid i,\theta^{(t)},K)\,P_r(p_n\mid i,\theta^{(t)},K)} \tag{12}$$

最大化(M 步)如式(13)所示,极大化 E 步求得后验概率更新超参数的值,并用于下一次迭代的 E 步。

$$T_n=\underset{k}{\operatorname{argmax}}\,\gamma(s_n) \tag{13}$$

再整合第 k 个子市场,根据房屋所属子市场,对 HPM 回归更新,即更新 $f_k^{(t+1)}$。计算出有效成员数量 $N_k^{(t+1)}=\sum_{n=1}^{N}\gamma(s_n)$,并最大化 $Q(\theta\mid\hat{\theta}^{(t)})$,令其求导为零,求得超参数 θ。具体的超参数值如式(14)所示:

$$a_k^{(t+1)}=\frac{N_k^{(t+1)}}{N}$$

$$\mu_k^{(t+1)}=\frac{1}{N_k^{(t+1)}}\sum_{n=1}^{N}\gamma(s_n)\,p_n$$

$$\sum_k^{(t+1)}=\frac{1}{N_k^{(t+1)}}\sum_{n=1}^{N}\gamma(s_n)\,(p_n-\mu_k^{(t+1)})\,(p_n-\mu_k^{(t+1)})^{\mathrm{T}}$$

$$\sigma_{y,k}^{2(t+1)}=\frac{1}{N_k^{(t+1)}}\sum_{n=1}^{N}\gamma(s_n)\,(y_n-f_k^{(t+1)}(r_n))^2 \tag{14}$$

当连续两次迭代得到的对数似然值之差小于阈值时,EM 算法迭代终止。

4.4 子市场划分与价格预测

求解出最优超参数后,使用贝叶斯平均值作为最终的预测价格。利用 EM 算法求得的参数 θ,计算房屋在每个子市场的后验概率,如式(15)所示:

$$P_r(s=k\mid p,r,\hat{\theta})=\frac{P_r(s=k\mid\hat{\theta})\,P_r(p,r\mid s=k,\hat{\theta})}{\sum_{i=1}^{K}P_r(s=i\mid\hat{\theta})\,P_r(p,r\mid s=i,\hat{\theta})} \tag{15}$$

求出该房屋在 k 个 HPM 中 k 个子市场的预测价格平均值,如式(16)所示,即作为最终预测价格:

$$\hat{y}=\sum_{k=1}^{K}\hat{y}_k P_r(s=k\mid p,r,\hat{\theta}) \tag{16}$$

再通过该房屋的后验概率值,导出其所属子市场的标签,如式(17)所示:

$$T=\underset{k}{\operatorname{argmax}}\,P_r(s=k\mid p,r,\hat{\theta}) \tag{17}$$

5 实验分析

5.1 数据获取

本文使用杭州房地产市场数据来评估模型,实验数据集的统计数据如表 2 所示。该数据来自于中国最大的在线房地产交易网络之一的 Lianjia.com,包括房价、8 种房屋特征以及 20 种 POI 信息[13-14]。

表 2 实验数据集的统计

Table 2 Statistics of experimental datasets

数据集	属 性	计 数
房屋数据	房屋数量	53 162
	房屋特征数量	9
交通设施数据	地铁站数量	257
	公交站数量	19 648
	停车场数量	36 165
中小学数据	公立普通学校数量	716
	公立重点学校数量	75
	私立普通学校数量	13
	私立重点学校数量	7
POI 数据	类型数量	20
	POI 总数	676 542

将房地产数据分为训练集、验证集和测试集。随

机抽选 60%的数据为训练集,20%数据作为验证集,剩下 20%数据作为测试集。

其中,9 种房屋特征包括年份、面积、楼层、朝向、装修度、价格、卧室数量、客厅数量、浴室数量。20 种 POI 信息包括餐饮服务类型、零售服务类型、专业市场类型、体育服务类型、娱乐健康类型、商务办公类型、产业园区、医疗卫生类型、文化艺术类型、教育科研类型、风景名胜、住宿服务类型、室内设施、金融保险类型、非制造业公司、制造加工类型、农业生成类型、公共管理类型、社会福利类型、其他生活服务类型等[15]。

5.2　数据预处理

数据清洗:所采集的房屋数据,存在部分基础特征缺失、环境特征缺失、价格存在离群值等问题。对特征缺失值采取丢弃、GAN 数据补充等方法,对于置信区间 0.95 外的离群值采取丢弃方法。

数据集成:将采集并清洗后的房屋基础特征数据和 POI 信息数据,根据“小区名”关键字,进行拼接整合到一起。

数据转换:将房屋基础特征中的年份、面积、楼层、朝向、装修度、卧室数量、客厅数量、浴室数量、20 种 POI 信息、交通设施数量、中小学数量等,进行独热编码。以方便后续的归一化和预测运算。

以下为本文生成器参数设置。输入维度(input_dim):40 维,隐藏层 1:128 个神经元,批标准化 1:128 个特征,隐藏层 2:256 个神经元,批标准化 2:256 个特征,隐藏层 3:512 个神经元,批标准化 3:512 个特征,隐藏层 4:1 024 个神经元,批标准化 4:1 024 个特征,输出层:40 维。激活函数均为 LeakyReLU,斜率为 0.2。

5.3　算法性能评估标准

本文选取平均绝对百分比误差(MAPD)作为主要指标评估房价预测的准确性,将原始误差标准化,消除数值过大存在的偏差。平均绝对误差(MAE)提供直观的误差度量,反映预测价格与实际价格的误差大小。均方根误差(RMSE)反应数据集的离散程度。定义如式(18)所示:

$$\varepsilon_{\mathrm{MAPD}} = \frac{100\%}{n}\sum_{j=1}^{n}\left|\frac{\hat{y}_j - y_j}{y_j}\right|$$

$$\varepsilon_{\mathrm{MAE}} = \frac{1}{n}\sum_{j=1}^{n}|\hat{y}_j - y_j| \tag{18}$$

$$\varepsilon_{\mathrm{RMSE}} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{y}_j - y_j)^2}$$

为了证明本文模型的性能,选用 5 种对比模型进行估值比较,评估结果如表 3 所示。

表 3　性能评估对比

Table 3　Performance assessment comparison　%

模　型	$\varepsilon_{\mathrm{MAPD}}$	$\varepsilon_{\mathrm{MAE}}$	$\varepsilon_{\mathrm{RMSE}}$
SVM	31.63	9 848	14 022
ANN	24.30	6 901	9 878
OLS	23.52	7 606	10 563
GBDT	18.99	5 322	7 707
LightGBM	12.40	4 024	6 221
BPM	11.07	3 347	5 701
GAN-BPM	9.64	3 048	4 264

5.4　对比模型结果分析

支持向量机(Support Vector Machine,SVM)是一种常用的二分类模型,通过将实例的特征向量映射到空间中的点来进行分类。其目标是找到最佳的分隔线或超平面,以有效区分两个类别的点。SVM 适用于不同规模和类型的数据集。它通过优化分类间隔来进行分类,特别擅长处理非线性问题和高维数据。

人工神经网络(Artificial Neural Network,ANN)是一种复杂的网络结构,由大量处理单元(神经元)相互连接而成,模拟了人脑组织的结构和运行机制。人工神经网络可以分为多层和单层两种结构。每一层都包含多个神经元,并且神经元之间通过连接权重进行连接,形成线性或非线性的分类模型。具有自动学习的能力,通过学习过程可以自动发现有效的特征表示,并提供准确的预测结果。ANN 在许多领域得到了广泛应用,包括图像和语音识别、自然语言处理、预测分析和模式识别等。

正交最小二乘(Orthogonal Least Squares,OLS)是一种特定形式的最小二乘法,用于拟合数据和估计模型参数。它在统计学、数学建模和机器学习等领域中被广泛应用。核心思想是通过选择一组正交基函数,将原始数据投影到这组基函数所构成的子空间上。然后,利用最小二乘法来拟合投影后的数据,得到模型的参数估计。通常用于处理具有高维特征的数据集,其中特征之间可能存在相关性。通过使用正交基函数,可以降低特征之间的相关性,从而减少模型的复杂度,并提高对数据的拟合效果。

GBDT 是一种迭代的决策树算法,由多棵决策树组

成,所有树的结论累加起来做最终答案。可以发现多种有区分性的特征以及特征组合。

LightGBM(Light Gradient Boosting Machine)是一种基于 GBDT 的开源机器学习框架。它的设计目标是提供高效的训练速度和低内存占用,以应对大规模数据集和高维特征的挑战。与传统的 GBDT 相比,LightGBM 采用基于直方图的决策树算法和并行多线程支持等的优化技术,使得它在性能上有很大的优势。

如表 3 所示,BPM 为贝叶斯概率模型[16],GAN-BPM 为加入数据增强模块的贝叶斯概率模型。本文提出的基于 GAN 的数据增强技术和子市场划分的贝叶斯概率模型,性能优于上述 5 种对比模型,精度有较大提高。

本文 BPM 模型基于子市场划分建立,子市场划分结果如图 5 所示。BPM 根据地理位置对位置相邻的房屋进行聚类,子市场中心点结果如图 6 所示。地理位置相邻的房屋最终划分到一个子市场内,如图 7 所示。每种颜色的圆点代表一个子市场内部的房屋。

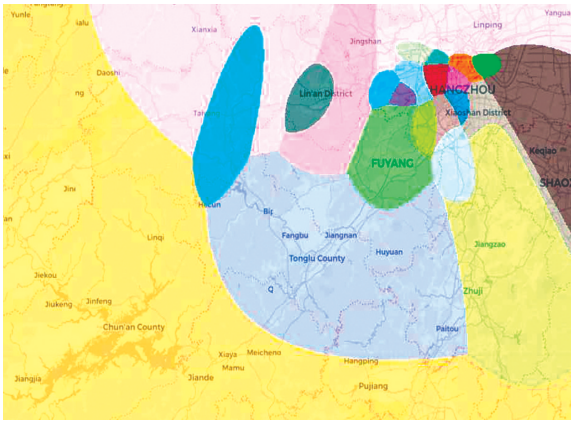


图 5 子市场划分结果

Fig. 5 Submarket division

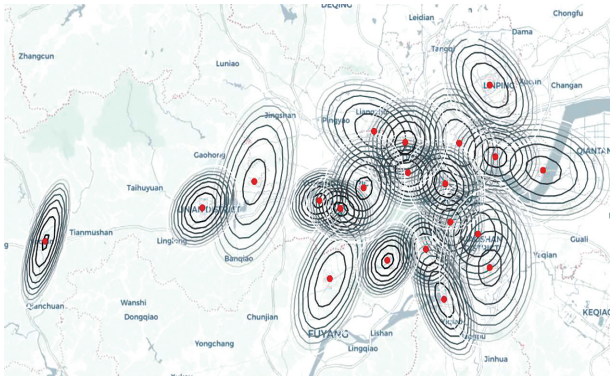


图 6 子市场中心点结果

Fig. 6 Submarket centers

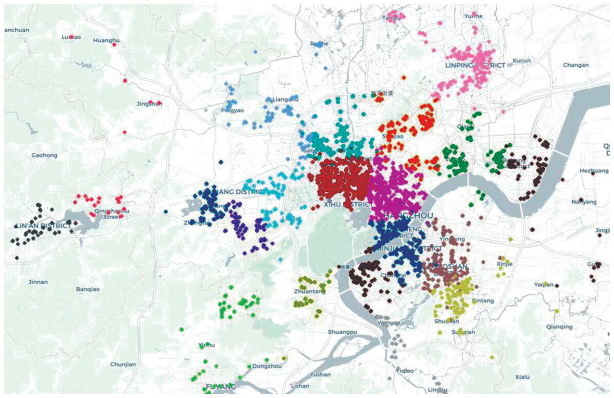


图 7 子市场房屋数量结果

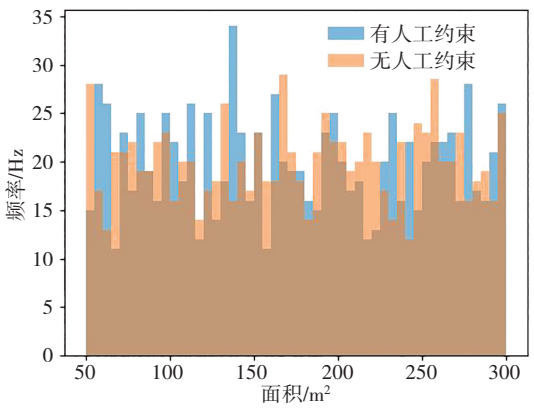
Fig. 7 Submarket housing quantity

5.5 人工约束对比

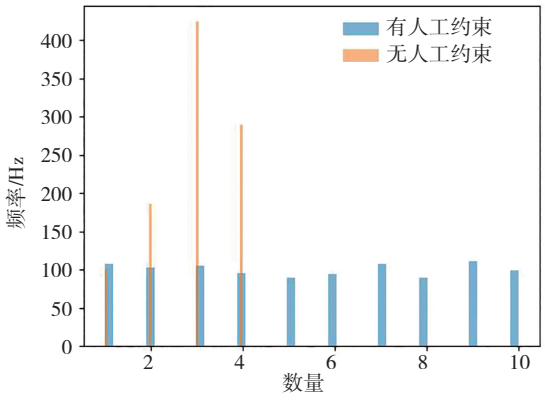
在无人工约束的情况下,生成模型完全依赖于数据集本身的分布特征进行训练和生成。这种方法的优点是生成过程自由度高,可以创造出更具创意和多样性的房屋数据。然而,缺点是生成的数据可能不符合实际的城市规划要求,如房屋布局、各类型房间数量不合理等。这样的数据可能在实际应用中缺乏可行性和实用性。

通过加入人工约束,可以确保生成的房屋数据符合实际的住房规划要求。如可以约束生成数据的房间数量与面积分配,合理房屋布局等。优点是生成的数据更符合实际规划要求,具有更好的可行性和实用性。此外,加入人工约束还可以提供更高的控制能力,可以根据特定的需求和规划要求生成符合特定条件的房屋数据。

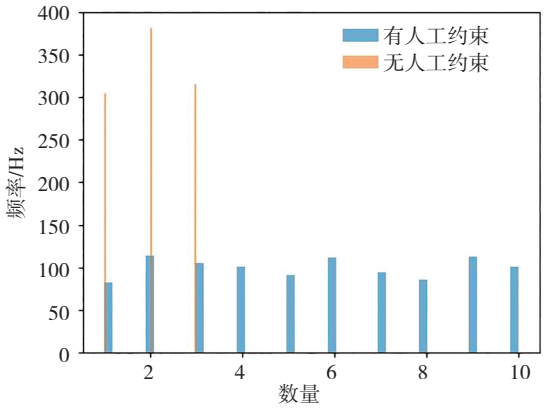
如图 8 所示,不在人工干预的情况下,房间数量、客厅数量、浴室数量等,会由于 GAN 生成器噪声生成,产生不符合实际的情况,在加入人工约束后,生成的数据,基本符合正常的住房规划需求。



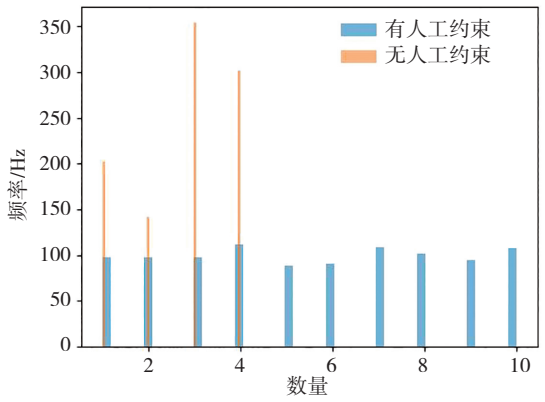
(a) 面积/m^2



(b) 卧室数量



(c) 客厅数量



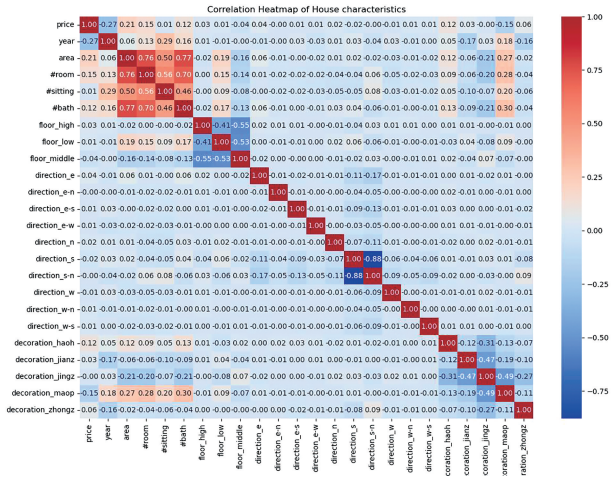
(d) 浴室数量

图 8　有无人工约束对比

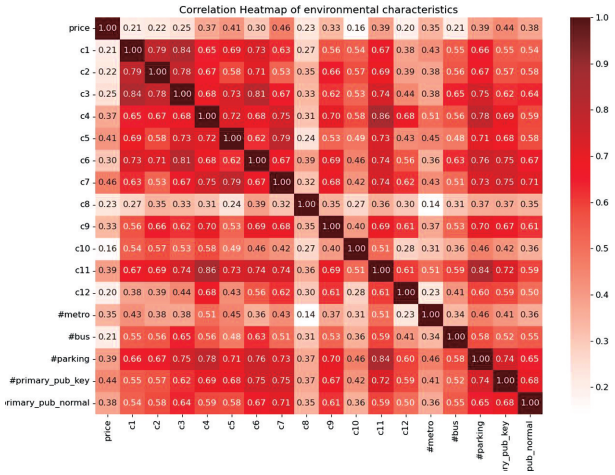
Fig. 8　Comparison with and without manual constraints

5.6　相关特征影响因素分析

如图 9 所示,根据生成的热力图,可以观察到高强度区域。热力图显示了明显的高强度区域,这些区域在颜色上呈现出深色或明亮的颜色,这表明在这些区域中的数据值较高。在本文的研究中,发现在房屋基础特征的情况下,房屋价格与面积、卧室以及浴室数量有较大关系。在环境特征中,娱乐康体职能、产业园区、风景名胜和周边学校数量能较大影响房屋的价格。



(a) 房屋价格基础特征热力图



(b) 房屋价格环境特征热力图

图 9　房屋价格特征热力图

Fig. 9　Heat map of house price characteristics

6　结　论

提出一种基于 GAN 的子市场划分房屋定价模型(GAN-BPM),旨在进行数据增强来提升子市场划分的效果。实验和分析表明 GAN-BPM 在数据增强方面表现出色。通过使用基于 GAN 的数据增强技术,成功地生成了具有多样性的合成数据,这些增强数据在训练机器学习模型时提升了原有模型性能;GAN-BPM 能够保留原始数据的特征和分布,并且能够生成更具挑战性的样本;GAN-BPM 在子市场划分中具有优势,通过贝叶斯概率模型的子市场划分方法,成功地对复杂市场数据进行划分,并识别出具有相似特征和需求的子市场;将 GAN-BPM 的数据增强和子市场划分方法相结合,实现了一个数据分析和应用框架。该框架不仅提供了增强的数据集,而且提供了对市场更细粒度的理解和洞察。

参考文献(References):

[1] SANDEEP KUMAR E, TALASILA V, PASUMARTHY R. A novel architecture to identify locations for Real Estate Investment[J]. International Journal of Information Management, 2021, 56: 102012.

[2] SOARES SILVA J C, DE ALMEIDA FILHO A T. Performing hierarchical Bayesian regression to assess the best districts for building new residential real estate developments[C]// Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. Piscataway: IEEE Press, 2020: 2411-2416.

[3] 袁韶华, 汪应宏, 左晓宝, 等. 徐州房地产空间异质分析及对城市规划的启示[J]. 工程经济, 2017, 27(12): 46-51.
YUAN Shao-hua, WANG Ying-hong, ZUO Xiao-bao, et al. Spatial heterogeneity analysis of real estate in Xuzhou and the inspiration for urban planning[J]. Engineering Economy, 2017, 27(12): 46-51.

[4] ZHANG Z, GAO Q, LIU L, et al. A high-quality rice leaf disease image data augmentation method based on a dual GAN[J]. IEEE Access, 2023, 11: 21176-21191.

[5] SUN Y. Real estate evaluation model based on genetic algorithm optimized neural network[J]. Data Science Journal, 2019, 18: 36-40.

[6] XU X, HUANG Z, WU J, et al. Finding the key influences on the house price by finite mixture model based on the real estate data in Changchun[C]//International Conference on Database Systems for Advanced Applications. Cham: Springer, 2019: 378-382.

[7] WU C, SHARMA R. Housing submarket classification: the role of spatial contiguity[J]. Applied Geography, 2012, 32(2): 746-756.

[8] LIU Z, CAO J, XIE R, et al. Modeling submarket effect for real estate hedonic valuation: a probabilistic approach[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(7): 2943-2955.

[9] 秦心静, 章平, 张新杨. 基于位置子市场划分的房价贝叶斯概率模型[J]. 重庆工商大学学报(自然科学版), 2023, 40(5): 81-88.
QIN Xin-jing, ZHANG Ping, ZHANG Xin-yang. Bayesian probability model for real estate price based on location submarket segmentation[J]. Journal of Chongqing Technology and Business University(Natural Science Edition), 2023, 40(5): 81-88.

[10] WU D, ZHANG Y. A Data Enhancement Method for Gene Expression Profile Based on Improved conditional SN-GAN[C]//Proceedings of the 2022 4th International Conference on Robotics, Intelligent Control and Artificial Intelligence. New York: ACM, 2022: 1059-1064.

[11] DOU Y, YU X, LI J. Feature GANs: A model for data enhancement and sample balance of foreign object detection in high voltage transmission lines[C]//International Conference on Computer Analysis of Images and Patterns. Cham: Springer, 2019: 568-580.

[12] HSIEH C F, LIN T C. Housing price prediction by using generative adversarial networks [C]//Proceedings of the International Conference on Technologies and Applications of Artificial Intelligence. Piscataway: IEEE Press, 2021: 49-53.

[13] 钟海玥, 张安录, 蔡银莺. 武汉市南湖景观对周边住宅价值的影响: 基于Hedonic模型的实证研究[J]. 中国土地科学, 2009, 23(12): 63-68.
ZHONG Hai-yue, ZHANG An-lu, CAI Yin-ying. Impacts of the Nanhu Lake in Wuhan city on the price of peripheral houses: Empirical research based on Hedonic model[J]. China Land Science, 2009, 23(12): 63-68.

[14] FU Y, XIONG H. Modeling of geographic dependencies for real estate ranking on site selection[C]//Proceedings of the 2015 IEEE International Conference on Data Mining Workshop. Piscataway: IEEE Press, 2015: 1506-1513.

[15] GOODMAN A C, THIBODEAU T G. The spatial proximity of metropolitan area housing submarkets [J]. Real Estate Economics, 2007, 35(2): 209-232.

[16] 李勇. 基于先验的贝叶斯先验选择方法[J]. 重庆工商大学学报(自然科学版), 2006, 23(6): 548-550.
LI Yong. On Bayesian Prior selection method based on prior[J]. Journal of Chongqing Technology and Business University(Natural Science Edition), 2006, 23(6): 548-550.

责任编辑:李翠薇