

改进 NRS 与 ELM 相结合在住宅需求预测中的应用^①



黄旭东¹, 狄晓涛², 沈明威¹

¹(河海大学 计算机与信息学院, 南京 211106)

²(南京国图信息产业有限公司, 南京 210036)

通信作者: 黄旭东, E-mail: hxd15056677631@163.com

摘要: 针对住宅需求预测受到不同方面因素的影响且具有非线性特征等问题, 本文在原始邻域粗糙集 (NRS) 的基础上进行改进, 并与极限学习机 (ELM) 相结合来进行预测. 首先改进算法 (MNRS) 解决了原始 NRS 无法在不同条件属性之间设定最佳邻域值的问题, 根据不同条件属性的邻域半径和标准差构建邻域关系矩阵; 然后在输出属性重要度排序时引入 Pearson 相关系数, 克服了条件属性之间的影响, 获得最小冗余属性的约简集构成住宅需求预测指标体系; 最后将构建的住宅需求指标体系输入极限学习机模型, 得到准确的预测值. 实验结果表明: MNRS-ELM 预测模型不仅有效降低了运算复杂度, 而且能够获得更高的预测精度.

关键词: 需求预测; 邻域粗糙集; 预测指标体系; 极限学习机

引用格式: 黄旭东, 狄晓涛, 沈明威. 改进 NRS 与 ELM 相结合在住宅需求预测中的应用. 计算机系统应用, 2024, 33(4): 302-307. <http://www.c-s-a.org.cn/1003-3254/9403.html>

Application of Improved NRS Combined with ELM in Residential Demand Prediction

HUANG Xu-Dong¹, DI Xiao-Tao², SHEN Ming-Wei¹

¹(College of Computer and Information Engineering, Hohai University, Nanjing 211106, China)

²(Nanjing Guotu Information Industry Co. Ltd., Nanjing 210036, China)

Abstract: Residential demand forecasting is affected by multiple factors and is non-linear. To address this issue, the study modifies the original neighborhood rough set (NRS) and then combines it with extreme learning machines (ELMs) to forecast residential demands. Specifically, the modified NRS (MNRS) algorithm constructs a neighborhood relationship matrix based on the neighborhood radii and standard deviations of different conditional attributes, thereby overcoming the failure of the original NRS algorithm to set the optimal neighborhood value for different conditional attributes. Then, the Pearson correlation coefficient is introduced into output attribute importance ranking to overcome the influence among conditional attributes, and the minimal redundant attribute-based reduction set is obtained to serve as the indicator system for residential demand forecasting. Finally, the residential demand indicator system is input into the ELM model to output an accurate forecasted value. Experimental results show that the MNRS-ELM forecasting model not only effectively reduces the operational complexity but also achieves higher prediction accuracy.

Key words: demand prediction; neighborhood rough set (NRS); predictive indicator system; extreme learning machine (ELM)

① 基金项目: 江苏省自然科学基金 (BK20221499)

收稿时间: 2023-08-24; 修改时间: 2023-09-26; 采用时间: 2023-10-08; csa 在线出版时间: 2024-03-04

CNKI 网络首发时间: 2024-03-08

1 引言

近年来,随着人口增长、城市化进程的加速和经济发展的推进,住宅需求的预测成为城市规划和房地产市场发展的重要问题.通过对住宅需求趋势的研究和分析,可以预测未来房地产市场的供求关系,从而使政府能够制定合理的住房政策、使房地产开发商能够确定合适的房屋建设计划、使个人投资者对投资房地产市场做出更明智的决策.因此,住宅需求预测对于促进经济发展、提高城市管理水平和改善人民生活质量具有重要意义.

需求预测主要分为两个方面:需求影响因素指标的选取和预测方法的选择.在影响因素指标选取方面,主要方法有主成分分析法^[1]、灰色关联分析法^[2]、邻域粗糙集法^[3]等.将因素指标通过特征提取之后输入构建好的算法预测模型获得住宅需求预测.而在预测方法选择方面,主要采用方法有两种:(1)以计量统计学为基础的传统预测方法,主要有多元线性回归分析模型^[4]、灰色系统模型^[5]、岭回归模型^[6]、时间序列分析模型^[7]与马尔可夫链模型^[8];(2)以机器学习和神经网络为基础的智能预测方法,主要有随机森林模型^[9]、SVM模型^[10]、BP神经网络模型^[11]、RBF神经网络模型^[12]、LSSVR模型^[13]等.

常见需求因素指标提取方法存在特征提取不充分的问题,而目前所使用的需求预测方法中存在不适用于非线性特征的情况、学习收敛速度慢等问题.针对此类问题,本文在原始邻域粗糙集(NRS)的基础上进行改进,并与极限学习机(ELM)相结合来预测住宅需求.首先会根据条件属性的邻域半径与标准差建立新的邻域关系矩阵,并将Pearson相关系数引入到邻域粗糙集来构建住宅预测指标体系;然后将改进的邻域粗糙集(MNRS)与极限学习机(ELM)相结合,建立基于MNRS-ELM的住宅需求预测模型;最后将MNRS-ELM与NRS-ELM以及ELM进行对比分析,实验结果表明,MNRS-ELM能够获得更高的预测精度.

2 最小冗余属性的邻域粗糙集算法

2.1 邻域粗糙集

定义1. 给定的 N 维的实数空间 Ω 里, $\Delta = R^N \times R^N \rightarrow R$,则称 Δ 是 R^N 上的一个度量^[14].在 N 维实数空间 Ω 上的非空有限集合 $U = \{x_1, x_2, \dots, x_N\}$ 中,对 $\forall x_i$ 的邻域 δ 定义为:

$$\delta(x_i) = \{x | x \in U, \Delta(x, x_i) \leq \delta\} \quad (1)$$

其中, $\delta \geq 0$, Δ 为距离函数, $\delta(x_i)$ 为 x_i 的邻域粒子.

定义2. 设条件属性为影响因素指标集合 C ,决策属性为住宅需求集合 D ,对于住宅需求邻域决策系统 $N = (U, C \cup D, V)$, V 是属性值的集合.对 $B \subseteq C$, $X \subseteq U$,邻域的上、下近似分别定义为:

$$\bar{N}_B(X) = \{x_i | \delta_B(x_i) \cap X \neq \emptyset, x_i \in U\} \quad (2)$$

$$\underline{N}_B(X) = \{x_i | \delta_B(x_i) \cap X, x_i \in U\} \quad (3)$$

其中,决策系统的正域 $Pos(X) = \underline{N}_B(X)$.

定义3. 在邻域决策系统中,存在指标 $a \in B$,那么 a 对于 D 的重要度为:

$$Sig(a, B, D) = \gamma_B(D) - \gamma_{B-(a)}(D) \quad (4)$$

其中, $\gamma_B(D)$ 是住宅需求 D 对因素指标子集 B 的依赖度,定义为:

$$k_D = \gamma_B(D) = \frac{|Pos_B(D)|}{|U|} \quad (5)$$

2.2 改进的构建邻域关系矩阵方法

原始的邻域粗糙集通常是在各自实验和专家判断经验的基础上,设置相同的邻域值给所有的因素指标.但在住宅需求因素指标的提取中,由于各因素指标的数据之间存在较大差异,邻域值设置不佳常常会导致因素指标约简效果较差^[15].因此本文根据不同因素指标的邻域半径和标准差为不同因素指标设置不同的邻域值,提出一种改进的邻域关系矩阵构建方法.

在住宅需求邻域决策系统中,邻域半径计算公式^[16]为:

$$\delta(a_i) = Std_{a_i} / \lambda \quad (6)$$

其中, Std_{a_i} 为因素指标 a_i 的标准差, λ 为邻域半径参数.

定义4. 住宅需求邻域决策系统 $N = (U, C \cup D, V)$,令该系统因素指标的标准差集合为 $\{Std_{a_1}, Std_{a_2}, \dots, Std_{a_n}\}$,定义 U 上第 a_i 个因素指标的邻域关系 N_{a_i} 的关系矩阵为:

$$M(N_{a_i}) = (r_{p,q})_{n \times n} \quad (7)$$

其中, $(r_{p,q})_{n \times n} = \begin{cases} 1, \Delta(x_p, x_q) \leq \delta(a_i) \\ 0, \text{其他} \end{cases}, 1 \leq p \leq n, 1 \leq q \leq n, \delta(a_i) = Std_{a_i} / \lambda, \lambda \in (0, 2]$.

2.3 基于Pearson相关系数的属性约简

一般计算属性重要度,只考虑单个条件属性对决

策属性的影响,并未将条件属性之间的影响考虑其中.若两个条件属性之间的关联性非常强烈,会导致最终输出的约简集数据冗余^[17],所以本文引入 Pearson 相关系数进一步剔除冗余属性,用来消除因素指标之间关联性的影响.

定义 5. 住宅需求邻域决策系统 $N = (U, C \cup D, V)$, $\forall a_i, a_j \in C$, 将 a_i 和 a_j 中的数据按照属性重要度从大到小分别进行排序,第 k 个样本在 a_i, a_j 指标下分别对应为 x_k 和 y_k , 则 a_i 和 a_j 的相关系数 ρ_{ij} 定义为:

$$\rho_{ij} = \frac{\left| \sum_{k=1}^{|U|} [(x_k - \bar{x})(y_k - \bar{y})] \right|}{\sqrt{\sum_{k=1}^{|U|} (x_k - \bar{x})^2 \sum_{k=1}^{|U|} (y_k - \bar{y})^2}} \quad (8)$$

其中, $\bar{x} = \frac{1}{|U|} \sum_{k=1}^{|U|} x_k$, $\bar{y} = \frac{1}{|U|} \sum_{k=1}^{|U|} y_k$. $0 \leq \rho_{ij} \leq 1$, ρ_{ij} 越大,表示指标 a_i, a_j 的相关性越高; ρ_{ij} 越小,表示指标 a_i, a_j 的相关性越低.

基于 Pearson 相关系数的属性约简,首先需计算全部因素指标的重要度,在约简集中加入重要度最大的因素指标.接下来要进行因素指标间的阈值判断.若 $\rho_{ij} < \eta$,那么在约简集中加入该因素指标.若 $\rho_{ij} \geq \eta$,那么先删除待加入的因素指标,然后按照重要度由高到低的顺序继续遍历剩余因素指标,直至结束.

一般来说,如果 Pearson 相关系数超过 0.8,表示因素指标之间的关联性非常强烈;如果 Pearson 相关系数在 0.6-0.8 之间,表示因素指标之间存在较大的关联性;如果 Pearson 相关系数在 0.4-0.6 之间,表示因素指标之间的关联性一般;如果 Pearson 相关系数低于 0.4,表示因素指标之间的关联性较弱^[17].所以本文设置相关系数阈值为 $\eta = 0.6$.

2.4 MNRS 算法步骤

MNRS 算法步骤的基本逻辑顺序为:首先根据指标的标准差和邻域半径构建邻域关系矩阵,然后根据指标重要度进行排序筛选,最后比较指标间的皮尔森相关系数约简输出约简集.具体步骤如算法 1 所示.

算法 1. MNRS 算法

输入: 邻域决策系统 $N=(U, C \cup D, V)$, 邻域半径参数 λ .

输出: 决策系统属性约简集 red .

step 1. 初始化 $red = \emptyset$, 样本 $smp = U$;

step 2. $\forall a_i \in C$, 对因素指标进行邻域划分, 计算得到邻域关系矩阵 N_{a_i} ;

step 3. $\forall a_i \in C - red$, 根据邻域关系矩阵 N_{a_i} 计算各个因素指标的重要度 $Sig(a_i, red, D)$;

step 4. 计算得到属性重要度最大的属性 a_j , 计算 a_i 和 a_j 的相关系数 ρ_{ij} , 并与阈值 η 比较, 去掉与 a_j 相关性大的属性, $red = red \cup a_j$;

step 5. 直到 $Sig(a_i, red, D) = 0$, 否则返回 step 3;

step 6. 输出决策系统约简集 red , 算法结束.

3 基于 MNRS 的 ELM 预测模型

3.1 ELM 预测模型

ELM 是一种新型的单隐层前馈神经网络^[18,19]. ELM 网络结构如图 1 所示.

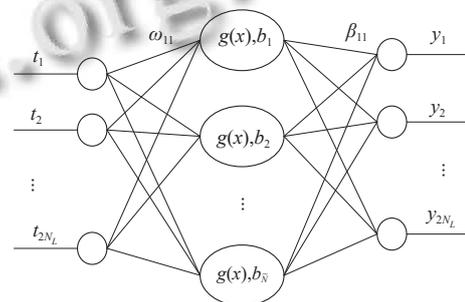


图 1 ELM 网络结构

在本文中,从第 2 节约简的指标体系中选取前 13 年的数据构建训练样本 $(t_i, y_i) (i = 1, 2, \dots, 13)$, ELM 的预测模型中有 \tilde{N} 个隐含层节点且激活函数为 $g(x)$ 的公式如下:

$$o_i = \sum_{j=1}^{\tilde{N}} \beta_j g(\omega_j \cdot t_i + b_j), \quad i = 1, 2, \dots, 13 \quad (9)$$

其中, $\beta_j = [\beta_{j1}, \beta_{j2}, \dots, \beta_{jm}]^T$ 为第 j 个隐含层和输出层的权重, $\omega_j = [\omega_{j1}, \omega_{j2}, \dots, \omega_{jm}]^T$ 为第 j 个隐含层和输入层的权重, b_j 为第 j 个隐含层节点的偏差, o_i 为第 i 个样本的输出值.

对于包含 \tilde{N} 个隐含层节点, 激活函数为 $g(x)$ 的 ELM 网络, 可以利用零误差来逼近训练样本:

$$\sum_{j=1}^{\tilde{N}} \|o_j - y_j\| = 0 \quad (10)$$

那么存在合适的 ω_j, b_j, β_j 使得:

$$y_i = \sum_{j=1}^{\tilde{N}} \beta_j g(\omega_j t_i + b_j), \quad i = 1, 2, \dots, 13 \quad (11)$$

式 (11) 可以简写为:

$$H\beta = Y \quad (12)$$

其中, H 为隐含层输出矩阵. 通过最小二乘法求解线性方程, 可得:

$$\hat{\beta} = H^+ Y \quad (13)$$

其中, H^+ 是 H 的摩尔-彭罗斯逆矩阵.

3.2 基于 MNRS 的 ELM 住宅需求预测流程

本文结合南京市住宅市场现状与 2005–2021 年统计年鉴数据进行住宅需求预测, 具体的流程图如图 2.

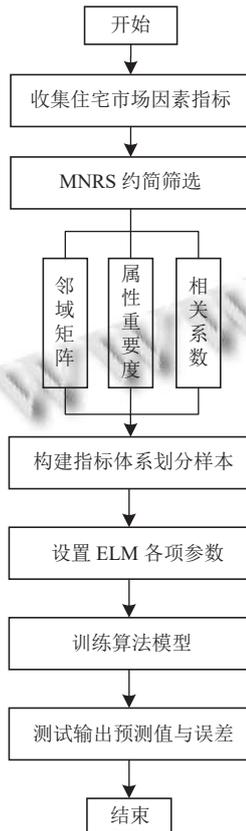


图 2 住宅需求预测流程

4 住宅需求预测实验分析

4.1 构建住宅预测指标体系

本文在南京市住宅市场的基础上, 结合房地产景气指标体系及中房指数指标体系, 选取了社会发展因素、家庭状况因素、住宅市场发展因素 3 个维度共 17 个影响因素为初选指标具体如表 1 所示. 住宅需求量以住宅销售面积来表示, 本文数据主要来源于南京市统计年鉴.

考虑到各因素指标之间存在量纲差异, 会导致训练时间过长、泛化能力较弱、出现较大误差. 所以在用 MNRS 构建指标体系之前, 先将数据归一化, 使数值控制在 $[0, 1]$ 的范围内.

表 1 影响因素初选指标

维度	序号	影响因素指标
社会发展因素指标	X_1	地区生产总值
	X_2	城镇化率
	X_3	人均GDP
家庭状况因素指标	X_4	家庭户数
	X_5	常住人口数
	X_6	人均可支配收入
	X_7	就业人数
住宅市场发展因素指标	X_8	消费价格指数
	X_9	人均居住消费支出
	X_{10}	人均居住面积
	X_{11}	房地产价格指数
	X_{12}	房地产开发投资额
	X_{13}	住宅开发投资额
	X_{14}	住宅施工面积
X_{15}	住宅竣工面积	
X_{16}	住宅销售均价	
X_{17}	住宅销售额	

将 17 个住宅需求影响因素初选指标分别经过 NRS 和 MNRS 筛选, 分别得到的约简后属性个数如表 2 所示.

表 2 约简后个数比较

数据样本	住宅预测数据个数
约简前	17
NRS约简后	9
MNRS约简后	7

根据 MNRS 算法步骤计算出因素指标重要度, 按照重要度进行相关系数的判定, 将相关系数大于 0.6 的指标剔除, 小于 0.6 的指标加入约简集. 由图 3 和表 2 可知, MNRS 算法约简后的个数与原始属性个数相比, 有了明显的减少. 故将约简后的因素指标 $X_3, X_8, X_{10}, X_{11}, X_{13}, X_{15}, X_{17}$ 构建住宅预测指标体系.

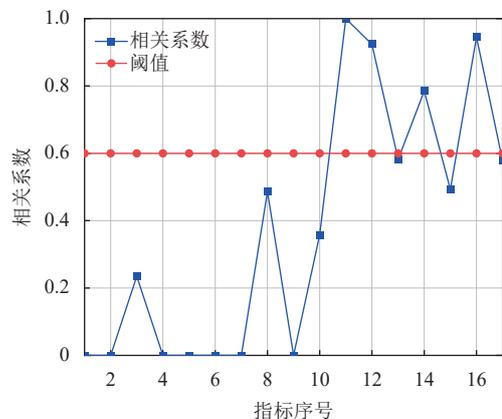


图 3 相关系数判定

构建住宅预测指标体系即为 ELM 算法的输入变量降维. ELM 的输入变量从原始的 17 个初选指标减少到指标体系中的 7 个指标, 最大程度地降低了算法复杂度与算法模型的运行时间.

4.2 ELM 参数设置

本文建立 3 层 ELM 神经网络, 内含输入层、隐含层和输出层, 利用 Matlab 进行 ELM 模型的搭建与训练. 分别将所有初选指标、NRS 约简的指标和 MNRS 约简的指标体系作为 ELM 模型的输入变量, 设置不同的输入节点个数, 然后将住宅需求作为输出变量, 设置输出节点个数为 1.

在 ELM 算法模型中, 输入层与隐含层的权重以及隐含层节点的偏差为随机设置^[20], 所以主要需要设置的参数为激活函数以及隐含层节点个数. 激活函数是运行在人工神经网络神经元上的函数, 其主要作用是向输出端映射神经元的输入值, 本文将激活函数设置为 Sigmoid 函数.

在 ELM 模型中, 当隐含层节点数目过少时, 网络无法拟合数据, 从而造成欠拟合状态; 当隐含层节点数目过多时, 不仅会增加训练时间, 还会降低学习效率和容错率, 进而影响模型的泛化能力. 所以本文将隐含层节点数目从 10 增加到 50, 由图 4 可知, 当隐含层节点数目为 30 时, 模型对训练样本的误差最小, 故本文 ELM 模型的隐含层节点个数设置为 30.

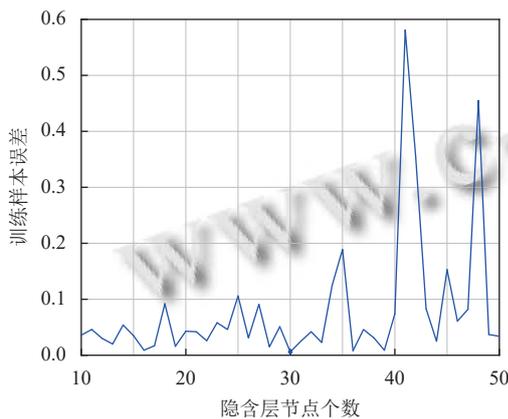


图 4 不同隐含层节点数目的误差

4.3 MNRS-ELM 模型预测分析

将约简后构成的住宅预测指标体系输入 ELM 预测算法模型, 选取 2005–2017 年的数据作为训练样本, 2018–2021 年的数据作为测试样本. 分别将全部因素指标、经邻域粗糙集约简后的指标因素以及经改进邻域

粗糙集约简后构建的住宅预测指标体系作为算法模型的输入, 住宅需求作为算法模型的输出, 结果如图 4.

由图 5 可知, 经 MNRS 约简后构建住宅预测指标体系的 ELM 模型预测值与真实值最为接近, 经 NRS 约简后因素指标的 ELM 模型预测值与真实值偏差中等, 全部因素指标的 ELM 模型预测值与真实值偏差较大, 表明 MNRS-ELM 的预测效果明显优于 ELM 的预测效果.

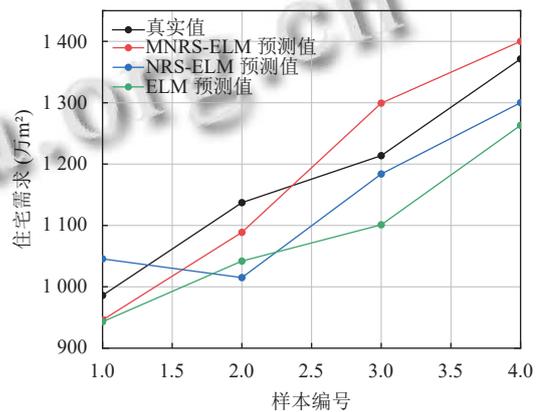


图 5 预测算法对比

同时本文还选择平均绝对百分比误差 (MAPE) 和拟合优度 (R^2) 作为评价住宅需求预测算法可靠性的主要评价指标, 并由此来对比分析算法模型的预测精度, 具体如表 3 所示.

表 3 预测精度比较

算法类别	MAPE (%)	R^2	时间 (s)
ELM	7.48	0.89	0.545
NRS-ELM	6.11	0.91	0.432
MNRS-ELM	4.37	0.95	0.423

由表 3 可知, MNRS-ELM 的平均绝对百分比误差比 ELM 降低了 3.1%, 且拟合优度提高了 6%, 另外运行效率也有所提升, 因此可以表明 MNRS-ELM 模型不仅有效降低了运算复杂度, 而且能够获得更高的预测精度.

5 结论

本文结合南京市住宅市场现状与 2005–2021 年统计年鉴数据, 将 MNRS-ELM 预测算法模型用于南京市住宅需求预测分析, 并将预测结果与 ELM 模型以及 NRS-ELM 模型进行对比, 得出以下结论.

- (1) 从社会发展因素、家庭状况因素、住宅市场

发展因素3个维度选取17个住宅预测影响因素初选指标,提出一种基于最小冗余属性的改进邻域粗糙集算法来约简影响因素指标,经MNRS约简后最终输出的7个因素指标构建了住宅预测指标体系。

(2)在住宅需求测试样本预测中,将经MNRS约简获得的住宅预测指标体系输入ELM算法模型,误差最小且预测值与真实值拟合效果最好,验证了MNRS-ELM预测模型的可行性。

参考文献

- 1 Kravchik M, Shabtai A. Efficient cyber attack detection in industrial control systems using lightweight neural networks and PCA. *IEEE Transactions on Dependable and Secure Computing*, 2022, 19(4): 2179–2197. [doi: [10.1109/TDSC.2021.3050101](https://doi.org/10.1109/TDSC.2021.3050101)]
- 2 Liu EC, Li J, Zheng AN, *et al.* Research on the prediction model of the used car price in view of the PSO-GRA-BP neural network. *Sustainability*, 2022, 14(15): 8993. [doi: [10.3390/SU14158993](https://doi.org/10.3390/SU14158993)]
- 3 Pan YZ, Xu WH, Ran QW. An incremental approach to feature selection using the weighted dominance-based neighborhood rough sets. *International Journal of Machine Learning and Cybernetics*, 2023, 14(4): 1217–1233. [doi: [10.1007/S13042-022-01695-4](https://doi.org/10.1007/S13042-022-01695-4)]
- 4 过娇雪. 哈尔滨市住房需求量预测研究 [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2013.
- 5 Xiong X, Guo H, Hu X. The housing demand analysis and prediction of the real estate based on the AWGM (1, N) model. *Grey Systems: Theory and Application*, 2021, 11(2): 222–240. [doi: [10.1108/gs-09-2019-0035](https://doi.org/10.1108/gs-09-2019-0035)]
- 6 纪甜甜. 房地产需求预测模型——以邯郸市住房需求为例 [硕士学位论文]. 邯郸: 河北工程大学, 2017.
- 7 顾昕雨. 基于ARIMA-SVR组合模型的卫星遥测数据预测研究 [硕士学位论文]. 北京: 中国科学院大学(中国科学院国家空间科学中心), 2021.
- 8 郑卓, 曹辉, 高鹤元, 等. 基于加权马尔可夫链修正的ARIMA预测模型的研究. *计算机应用与软件*, 2020, 37(12): 63–67.
- 9 张倩. 基于随机森林回归模型的住房租金预测模型的研究 [硕士学位论文]. 长春: 东北师范大学, 2019.
- 10 Wang XJ, Zeng GT, Zhang KX, *et al.* Urban real estate market early warning based on support vector machine: A case study of Beijing. *International Journal of Computational Intelligence Systems*, 2020, 13(1): 153–166. [doi: [10.2991/ijcis.d.200129.001](https://doi.org/10.2991/ijcis.d.200129.001)]
- 11 滕渊. 青岛市房地产市场预测模型研究 [硕士学位论文]. 青岛: 中国海洋大学, 2013.
- 12 王亚龙. 徐州市区城市建设用地需求预测研究 [硕士学位论文]. 徐州: 中国矿业大学, 2016.
- 13 Wang LD, Rong X, Chen ZY, *et al.* A real estate early warning system based on an improved PSO-LSSVR model—A Beijing case study. *Buildings*, 2022, 12(6): 706. [doi: [10.3390/BUILDINGS12060706](https://doi.org/10.3390/BUILDINGS12060706)]
- 14 胡清华, 于达仁, 谢宗霞. 基于邻域粒化和粗糙逼近的数值属性约简. *软件学报*, 2008, 19(3): 640–649.
- 15 毛清华, 马宏伟, 张旭辉. 改进邻域粗糙集的输送带缺陷特征约简算法. *仪器仪表学报*, 2014, 35(7): 1676–1680.
- 16 安若铭, 索明亮. 邻域粗糙集在属性约简及权重计算中的应用. *计算机工程与应用*, 2016, 52(7): 160–165.
- 17 林芷欣, 刘遵仁, 纪俊. 基于k近邻属性重要度和相关系数的属性约简. *计算机工程与设计*, 2020, 41(9): 2488–2494.
- 18 Huang GB, Zhu QY, Siew CK. Extreme learning machine: Theory and applications. *Neurocomputing*, 2006, 70(1-3): 489–501. [doi: [10.1016/j.neucom.2005.12.126](https://doi.org/10.1016/j.neucom.2005.12.126)]
- 19 Huang GB, Zhou HM, Ding XJ, *et al.* Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2012, 42(2): 513–529. [doi: [10.1109/TSMCB.2011.2168604](https://doi.org/10.1109/TSMCB.2011.2168604)]
- 20 骆公志, 侯若娴, 陈圣瑜. 基于邻域粗糙集的极限学习机恶意DoH流量预警. *南京邮电大学学报(自然科学版)*, 2022, 42(6): 79–85.

(校对责编: 孙君艳)