引用格式:邬嘉怡,王思玉,史宏炜,等. 基于多小波的北京市房屋市场价格的分析预测[J]. 北京化工大学学报(自然科学版),2019,46(5):101 - 106.

WU JiaYi, WANG SiYu, SHI HongWei, et al. A multi-wavelet transform for analysis of Beijing house prices[J]. Journal of Beijing University of Chemical Technology (Natural Science), 2019,46(5):101 - 106.

基于多小波的北京市房屋市场价格的分析预测

邬嘉怡 王思玉 史宏炜 李虎森 楼凯达 崔丽鸿*

(北京化工大学 理学院, 北京 100029)

摘 要: 针对房价波动大的数据特征,将多小波分析与房价预测问题结合,以北京市 2010—2018 年的房屋数据作为研究对象,探究了利用 Haar 小波变换、Daubechies 系列小波变换以及基于过采样预处理的 GHM 多小波变换和 CL 多小波变换处理房价数据的分解重构效果,并通过对高频系数进行门限阈值量化重构处理以达到去噪的目的;建立支持向量机(SVM)预测模型,通过探究 4 种小波处理方法对房屋价格预测的影响结果,给出了相应预测效果更佳的数据处理方法和选择依据。

关键词: 多小波变换; 离散小波变换(DWT); 软阈值去噪; 北京市房屋成交价格; 支持向量机(SVM)

中图分类号: O29 **DOI**: 10.13543/j.bhxbzr.2019.05.015

引 言

近年来,房价问题日渐升温,人们在关注房价问题的过程中,最关注的是房价的未来走势。但是,由于房价在历史时点上的数据波动巨大且具有信噪比低、信噪难以分离的特点,另外,影响房屋价格的不仅有时间,还有房屋面积、所处区域、房屋配置等指标,导致房屋指标与房价关系难以用传统预测方法构造,更难以给出有效的预测方法。因此如何高效处理房价数据使其适用于拟合和预测,具有重要的研究价值。

以往的研究主要立足于房价预测。杨楠等[1]采用灰色马尔可夫模型和 n 次多项式模型预测了全国房屋年平均价格;李佳音[2]提出市场比较法来预测房价;闫妍等[3]提出了基于 TEI@I 方法论的房价预测方法;Anglin[4]引入平均房价增长率及 CPI 等指标建立 VAR 模型来预测多伦多房价。但对于我国的房产市场,上述方法各有其适用范围和局限性。灰色马尔科夫模型只能预测短期趋势;基于 TEI@I 方法论的方法只适用于中短期预测;市场比较预测方法及国外模型只能比较成熟程度高、运作完善的房产市场,中国房产市场显然不具备类似条件。

有效的数据分析处理工具是探究我国房产市场发展规律和预测房价的关键。在诸多数据处理方法中,小波变换是一种信号的时间-尺度分析方法,它具有多分辨率分析的特点,能够在时、频两域较好地呈现信号的局部特征。基于小波函数的多尺度特性,可以将历史房价看作特定的信号,通过小波分析将其分解重构,再进行降噪处理,从而降低房价数据的非平稳性,使其能够运用传统预测模型来进行预测。但是除了 Haar 小波之外,现有研究常用的单小波不能同时满足正交、对称及紧支性(在实数范围内),而多小波可同时拥有这些应用上所需要的优良性质。因此本文提出基于多小波的方法,结合支持向量机预测模型,对北京市房屋市场价格进行分析预测。实验结果表明,相对于单小波,理论性质优越的多小波在应用上也表现出良好的特性。

1 基本理论

1.1 小波分析及其分解重构算法

多小波分析(multi wavelet analysis, MWA)是小波理论的新发展,单小波由一个母函数(基本函数)

收稿日期: 2018-11-20
基金项目: 国家级大学生创新训练计划(201810010050)
第一作者: 女,1996 年生,本科生
*通信联系人
E-mail: cuilh@mail.buct.edu.cn

通过伸缩平移得到的小波基构成,而多小波的基本母函数不止一个,因此其同时具有对称性、正交性、插值性、紧支性和高阶消失矩等特点,在理论上是优于单小波的一种数据分析方法。

小波分析的基本框架是多分辨率分析(MRA)[5]。当 $L^2(R)$ 空间一串闭子空间序列 $\{V_j\}_{j\in Z}$ 同时满足单调性($V_j \subset V_{j+1}$)、逼近性($\bigcap_{j\in Z} V_j = \{0\}$, $\overline{\bigcup_{j\in Z} V_j} = L^2(R)$)、伸缩性($f(t)\in V_j \Leftrightarrow f(2t)\in V_{j+1}$)、平移不变性($f(t)\in V_j \Leftrightarrow f(t-k)\in V_j$),并且存在函数 $g(t)\in V_0$,使得 $\{g(t-k)\}_{k\in Z}$ 构成空间的 Riesz 基时,称这个空间序列为多分辨率分析。

基于多分辨率分析的定义,$\varphi(t)\in V_0 \subset V_1$ 和 $\phi(t)\in W_0 \subset V_1$ 都可以用 V_1 空间的一个基 $\{\varphi(2t-k)\}_{k\in Z}$ 表示,即双尺度方程[6]

$$\begin{cases}\varphi(t) = \sum_k h_k \varphi(2t-k) \\ \psi(t) = \sum_k g_k \varphi(2t-k)\end{cases} \quad k\in Z \tag{1}$$

式中,$h_k = \langle \varphi(t), \varphi(2t-k)\rangle$,$g_k = \langle \psi(t), \varphi(2t-k)\rangle$。从信号分析的角度,$h$ 是与 φ 对应的低通滤波器,g 是与 ψ 对应的高通滤波器,$\{h,g\}$ 为滤波器组。

类似地,由 MRA 可以推出 r 重分辨率分析(MRA^r)的定义[7],构造相似的双尺度方程

$$\begin{cases}\boldsymbol{\Phi}(t) = \sum_k \boldsymbol{H}_k \boldsymbol{\Phi}(2t-k) \\ \boldsymbol{\Psi}(t) = \sum_k \boldsymbol{G}_k \boldsymbol{\Psi}(2t-k)\end{cases} \quad k\in Z \tag{2}$$

对任意的输入信号,有小波分解公式

$$\begin{cases} f(t) = \sum_{i=1}^{r}\sum_{k\in Z} c_{i,J,k}\varphi_{i,J,k}(t) = \\ \qquad \sum_{i=1}^{r}\sum_{k\in Z} c_{i,J_0,k}\varphi_{i,J_0,k}(t) + \\ \qquad \sum_{i=1}^{r}\sum_{J_0<j<J}\sum_{k\in Z} d_{i,j,k}\psi_{i,j,k}(t) \\ c_{i,J_0,k} = \int f(t)\varphi_{i,J_0,k}(t)\,\mathrm{d}t \\ d_{i,j,k} = \int f(t)\psi_{i,j,k}(t)\,\mathrm{d}t \end{cases} \tag{3}$$

基于式(1)可知,多分辨率分析的主要思想是将 $L^2(R)$ 分解为一串具有不同分辨率的子空间序列,将 $L^2(R)$ 中的函数 $f(t)$ 描述为具有一系列近似函数的逼近极限[8],其中每一个近似函数都是 $f(t)$ 在不同分辨率子空间上的投影,从而通过分析这些投影来获得近似函数的形态和特征。

本文将价格信号分成 5 层,其小波分解树示意图如图 1 所示。

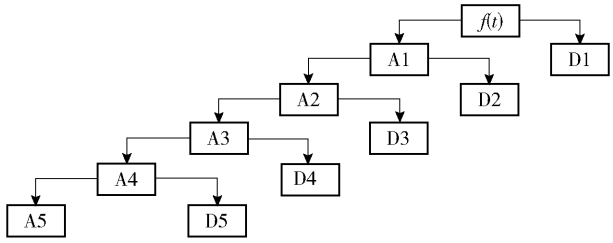

图 1 5 层小波分解树示意图

Fig. 1 A five-layer wavelet decomposition tree diagram

从图中可以看出,通过小波分解可得到逼近分量系数(低频部分)和细节分量系数(高频部分),其分解具有以下关系

$$f(t) = A_1 + D_1 + D_2 + D_3 + D_4 + D_5 \tag{4}$$

式中,A_1 为第一层分解的低频部分分量系数,D_i 为第 i 层分解的高频部分分量系数。

通常,有用信号表现为低频部分,噪声信号表现为高频部分。本文对小波分解的高频系数进行门限阈值量化处理,然后根据小波分解的第 5 层低频系数和经过量化后的 1 ~ 5 层高频系数进行小波重构,达到消除噪声的目的。由于本文的研究对象是价格变化,其在时间尺度下呈连续趋势,所以采用能够平滑化处理的软阈值进行量化去噪。

1.2 支持向量机

支持向量机(SVM)[9]是一种分类机器学习算法,其基本原理是利用核函数将输入样本空间映射到高维特征空间,然后在这个高维空间中求解最优分类面,得到输入与输出变量的非线性关系。

在 SVM 算法中,给定特征空间上的训练样本

$$\boldsymbol{T} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}, \boldsymbol{x}_i \in R^n, y_i \in R, i = 1, \cdots, n \tag{5}$$

式中,n 表示样本实例个数,$\boldsymbol{x}_i$ 表示第 i 个特征向量,y_i 为第 i 个预测值。

对于训练样本,存在一个分类面$(\boldsymbol{w}\cdot\boldsymbol{x}) + b = 0$,通过引入松弛变量 ξ_i,构建的最优分类面满足

$$y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i) + b) \geqslant 1 - \xi_i,\ \xi_i \geqslant 0,\ i = 1, 2, \cdots n \tag{6}$$

式中,$\boldsymbol{w}$ 为权值向量,b 为偏差项。

为了使预测值落入不同的分类面,要保证分类间隔最大,即目标函数 $O(\boldsymbol{w}, \xi_i)$ 值最小,此时最优分类面构造问题即转化为二次规划问题式(7)

$$\begin{cases} \min O(\boldsymbol{w},\xi_i) = \dfrac{1}{2}(\boldsymbol{w}\cdot\boldsymbol{w}) + C\sum_{i=1}^{n}\xi_i \\ \text{s.t. } y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b) \geqslant 1-\xi_i, \\ \quad \xi_i \geqslant 0, i = 1,2,\cdots,n \end{cases} \tag{7}$$

式中，C 是定义为常数变量的惩罚参数。

同时，引入核函数 $K(\boldsymbol{x},\boldsymbol{y}) = \phi(\boldsymbol{x})\phi(\boldsymbol{y})$，利用 Lagrange 乘子法以及 KKT 定理，将式(7)转化为对偶二次规划问题

$$\begin{cases} \max\ L(\boldsymbol{w},b,\alpha) = \sum_{i=1}^{n}\alpha_i - \\ \quad \dfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \\ \text{s.t. } \sum_{i=1}^{n}\alpha_i y_j = 0, \alpha_i \geqslant 0 \end{cases} \tag{8}$$

由式(8)得到非线性分类问题的判别函数

$$f(\boldsymbol{x}) = \operatorname{sgn}(\boldsymbol{w}\phi(\boldsymbol{x}) + b) = \operatorname{sgn}\left(\sum_{i=1}^{n}\alpha_i y_i K(\boldsymbol{x},\boldsymbol{y}) + b\right) \tag{9}$$

根据式(6)判别 $\boldsymbol{x}$ 的类别，即为预测值。

2　模型建立

2.1　数据处理及分析

收集北京市房屋市场 2010—2018 年的 421 891 条相关数据，包括每天每笔成交的单价、面积、户型、朝向、装修类别、电梯数、楼层、总层数、建造年代、房屋结构、所处区域等相关指标，删除其中有缺失值的记录。

由于收集的数据来自于每天成交网站，考虑到房价的时间成本，将单价以 0.7% 的贴现率按季度贴现成现值。处理后的房价数据原始状态散点图如图 2 所示。

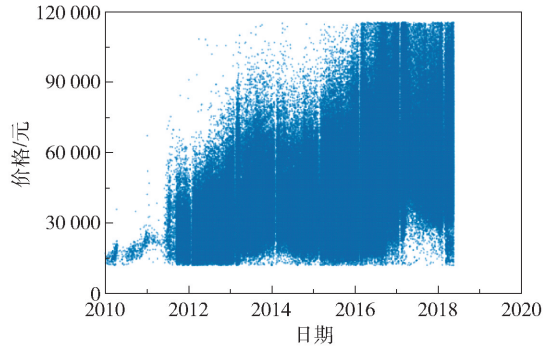

图 2　北京市房价与时间散点图

Fig. 2　Beijing house price and time scatter chart

由图 2 可以看出，在本预测模型中，房价数据信噪比低，信噪难以有效分离；且数据维数高、波动大，不能直接应用于预测模型。

利用小波变换可将任一时间段内的数据高频部分和低频部分分离，用高频部分反映房屋市场的短期变化趋势，低频部分反映中、长期变化趋势，使数据适用于传统预测模型。

2.2　小波基的选择

小波基可以用较少非零小波系数有效逼近实际函数，这一特性被广泛应用于数据压缩、信号去噪以及快速计算中，所以选择小波基应以最大量产生接近于零的小波系数为最优[10]。在小波分析的应用中，不同的小波基或小波函数的选取会产生不同的结果，要把握小波函数的特征，包括消失矩、正则性、紧支性、对称性以及正交性和双正交性等，根据应用的需要选择合适的小波基。

表 1 简要概括了常用小波基的特点[11]。本文基于小波基的特点选取了最简单的 Haar 小波基函数以及目前应用最广的 Daubechies(Db)系列小波进行研究。

表 1　常用小波基特点

Table 1　Common wavelet base features

小波基函数	消失矩	正交性	对称性	紧支性	支集宽度
Haar	1	有	有	有	1
Meyer	无	有	有	无	无穷
Morlet	1	无	有	无	无穷
Daubechies	N	有	无	有	$2N-1$
Symlet	N	有	近似	有	$2N-1$
coifN	$2N$	有	近似	有	$6N-1$

Haar 小波基函数是所有母函数中最简单的一种，也是唯一有对称和反对称的单小波，但 Haar 小波的消失矩为 1，对大于一次多项式的函数的消失效果不好。Db 小波基系列函数是基于消失矩构造的 p 阶消失矩的小波，同时具有良好的正则、正交和紧支性性质，因此应用十分广泛，本文选取 Db5 作为母函数。

2.3　多小波的选取

2.3.1　GHM 多小波

GHM 多小波是由 Geronimo 等[12]通过分形插值函数的方法给出的多小波系统，其支集长度为 4。GHM 多小波的尺度函数和小波函数都具有紧支性，其支集分别为[0,1]和[0,2]，因此具有良好的局域性；其尺度函数和小波函数具有对称性，尺度函数是

整数的平移正交,变换后能够保持能量恒定;同时系统存在二阶逼近。

2.3.2 CL 多小波

CL 多小波是 Chui 等[13]利用对称性给出的支集为[0,2]和[0,3]的多小波系统,包括 CL3 多小波(支集长度为 3)和 CL4 多小波(支集长度为 4),其中 CL3 多小波位于区间[0,2]上,CL4 多小波位于区间[0,3]上。CL 多小波的尺度函数和小波函数都具有紧支性,两个尺度函数分别与两个小波函数对称和反对称,保证了其线性相位;CL 多小波同时具有正交性;系统存在三阶逼近,其逼近性能优于 GHM 多小波。

2.4 核函数和参数的选择

以小波分析分解重构后的数据作为样本,建立 SVM 预测模型,预测后通过特征系数重构给出最终预测结果。

由于预测结果不能保证其线性,使用非线性 SVM 和核函数将变量映射到高维空间,选取了高斯核[14]

$$K(\boldsymbol{x},\boldsymbol{y}) = \exp\left(-\gamma|\boldsymbol{x}-\boldsymbol{y}|^2\right) \tag{10}$$

式中,核参数 $\gamma = \frac{1}{\sqrt{2}}$, SVM 预测模型式(7)中的惩罚参数 $C = 20$。

3 结果与讨论

对收集的房价数据进行小波去噪处理,选取其中一个区域的约 5 000 个数据进行降噪,比较 Haar 小波、Db 小波、GHM 多小波以及 CL 多小波的重构效果,然后用小波处理后的数据及其影响因子进行 SVM 房价预测,比较不同方法处理数据对预测结果的影响。

3.1 整体趋势

选取不同的单小波和多小波作为小波基对数据进行去噪,用 Matlab 编程,运行后分别得到基于 Haar 单小波、Db5 单小波、GHM 多小波及 CL 多小波软阈值去噪前后的散点对比图,如图 3 所示。

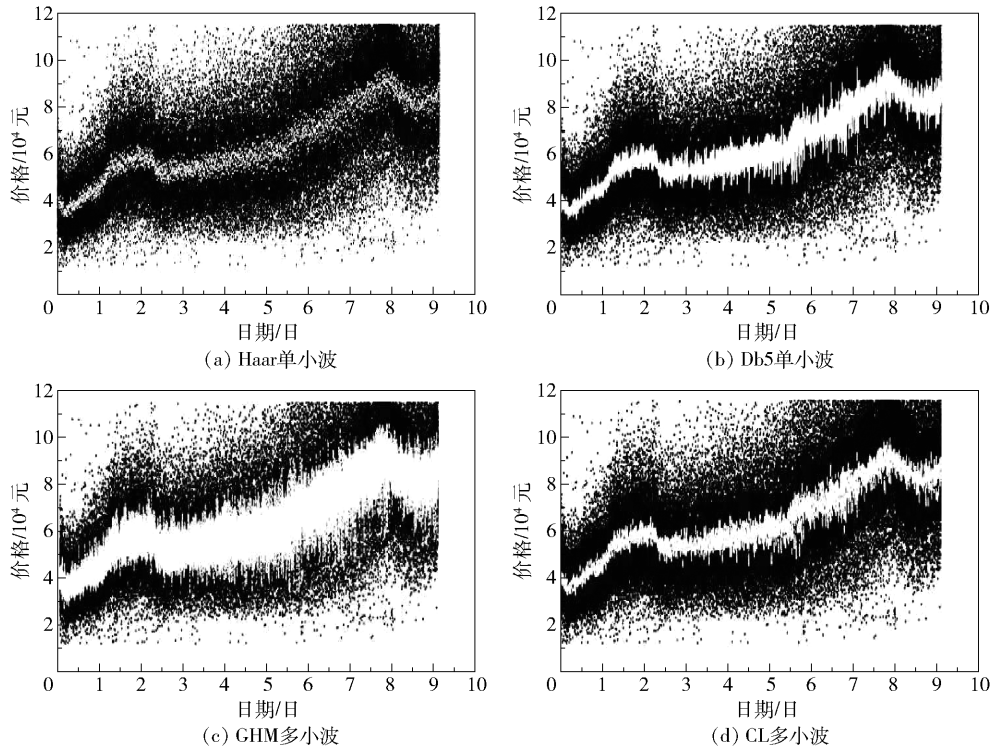

图 3 不同方法小波去噪前后数据散点对比

Fig. 3 Comparison of data scatter points before and after wavelet denoising using different methods

图 3 中黑、白色点分别为原始数据散点和去噪后散点。通过对比可以看到,无论是单小波还是多小波,去噪后数据的整体趋势与原始数据一致,说明小波去噪能保留数据的长期变化特征。

3.2 分解重构效果

为了说明单小波以及多小波分解重构对信号的影响,选取处理前后的数据标准误差、中位标准误差和平均标准误差对重构结果进行评价,结果如表 2

所示。

表 2　小波分析误差比较

Table 2　Comparison of wavelet analysis errors

小波类别	数据标准误差	中位标准误差	平均标准误差
Haar 单小波	4.679	3.277	3.891
Db5 单小波	4.679	3.437	3.864
GHM 多小波	4.265	3.167	3.763
CL 多小波	4.445	3.268	3.659

由表 2 数据综合比较看出，采用 GHM 多小波进行信号的分解、重构，能够较好地保留原始信号中的特征信息，且从该组数据来看，多小波的分解重构能力强于单小波。

3.3　房价预测效果

根据北京市房价的特点，将单小波（Haar、Db5）处理后数据以及多小波（GHM、CL）处理后数据相对应的 5 000 个样本代入 SVM 模型进行预测，并与原始数据直接预测的结果进行对比。将实际样本落入的等级称为“原始等级”，预测值落入“原始等级”且误差在 20% 区间内的预测结果可以接受。将落入可接受区间内的占比作为预测准确率，预测效果对比如表 3 所示。可以看出，用 CL 多小波处理后的数据预测准确率最高，预测效果最好，说明基于 CL 多小波的去噪处理能够相对最大程度地保留原始房价数据特征，且降低数据波动性，适合用于此类预测。

表 3　SVM 预测准确率比较

Table 3　Comparison of SVM prediction accuracy

数据来源	预测准确率/%
原始数据	70.62
Haar 单小波	87.2
Db5 单小波	84.26
GHM 多小波	88.9
CL 多小波	90.24

4　结论

（1）基于多小波的对称性、正交性、紧支性等优点，比较了以 Haar、Db5 为母函数的单小波分析，以及经过采样预处理的 GHM 和 CL 多小波分析的重构效果，证明小波去噪可以保留房价的变化趋势；通过重构误差分析发现多小波分析处理信号效果误差优于单小波，多小波分析更能保持原有信息的特征。

（2）SVM 模型房屋价格预测结果表明，CL 多小波分析处理后数据的预测结果准确率最高；在非平稳序列的预测中，小波分析处理数据能够优化传统预测结果，而多小波分析预测准确率高于单小波分析。

参考文献：

[1] 杨楠，邢力聪. 灰色马尔可夫模型在房价指数预测中的应用[J]. 统计与信息论坛，2006，21(5)：52-55.
YANG N，XING L C. Application of grey-Markov model on the prediction of housing price index[J]. Statistics & Information Forum，2006，21(5)：52-55.（in Chinese）

[2] 李佳音. 一种商品房价格预测方法[J]. 商品与质量，2011(9)：219-220.
LI J Y. A method for forecasting the price of commercial houses[J]. Trade and Quality，2011(9)：219-220.（in Chinese）

[3] 闫妍，徐伟，部慧，等. 基于 TEI@I 方法论的房价预测方法[J]. 系统工程理论与实践，2007，27(7)：1-9.
YAN Y，XU W，BU H，et al. House price forecasting method based on TEI@I methodology[J]. Systems Engineering Theory & Practice，2007，27(7)：1-9.（in Chinese）

[4] ANGLIN P. Local dynamics and contagion in real estate markets[C]// The International Conference on Real Estates and Macro Economy. Beijing，2006.

[5] 孙延奎. 小波分析及其应用[M]. 北京：机械工业出版社，2005：1-16.
SUN Y K. Wavelet analysis and its application[M]. Beijing：Mechanical Industry Press，2005：1-16.（in Chinese）

[6] 唐远炎，王玲. 小波分析与文本文字识别[M]. 北京：科学出版社，2004：44-53.
TANG Y Y，WANG L. Wavelet analysis and text recognition [M]. Beijing：Science Press，2004：44-53.（in Chinese）

[7] MALLAT S. Wavelet for a vision[J]. Proceedings of the IEEE，1996，84(4)：604-614.

[8] LEBRUN J，VETTERLI M. Balanced multi wavelets theory and design [J]. IEEE Transactions on Signal Processing，1998，46(4)：1119-1125.

[9] 丁世飞，齐丙娟，谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报，2011，40(1)：1-9.
DING S F，QI B J，TAN H Y. A survey of support vector machine theory and algorithms[J]. Journal of University of Electronic Science and Technology of China，2011，40(1)：1-9.（in Chinese）

[10] 李建平, 唐远炎. 小波分析方法的应用[M]. 重庆: 重庆大学出版社, 1998: 72-87.
LI J P, TANG Y Y. Application of wavelet analysis method [M]. Chongqing: Chongqing University Press, 1998: 72-87. (in Chinese)
[11] 高成. Matlab 小波分析与应用[M]. 2 版. 北京: 国防工业出版社, 2007: 27-28.
GAO C. Matlab wavelet analysis and application [M]. 2nd ed. Beijing: National Defense Industry Press, 2007: 27-28. (in Chinese)
[12] GERONIMO J S, HARDIN D P, MASSPOPUST P R. Fractal functions and wavelet expansions based on several scaling functions [J]. Journal of Approximation Theory, 1994, 78: 373-401.
[13] CHUI C K, LIAN J A. A study of orthonormal multi-wavelets [J]. Applied Numerical Mathematics, 1996, 20 (3): 273-298.
[14] 奉国和. SVM 分类核函数及参数选择比较 [J]. 计算机工程与应用, 2011, 47(3): 123-128.
FENG G H. SVM classification and function and parameter selection comparison [J]. Computer Engineering and Applications, 2011, 47(3): 123-128. (in Chinese)

A multi-wavelet transform for analysis of Beijing house prices

WU JiaYi WANG SiYu SHI HongWei LI HuSen LOU KaiDa CUI LiHong*
(Faculty of Science, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: In recent years, housing problems have become more prominent. However, due to the large fluctuations in data, it is difficult to predict the price of the housing market. By focusing on the data characteristics of house prices, this paper combines multi-wavelet analysis with house price forecasting. Using Beijing house price data, the effects of manipulating these data with Haar, Daubechies wavelet transform and GHM, and a CL multi-wavelet transform based on oversampling was compared. By means of a support vector machine (SVM), a model is established to predict unit-price based on treating the data with different wavelet transforms. Finally, a strategy for selecting house price data manipulation methods is suggested.

Key words: multi-wavelet transform; discrete wavelet transform (DWT); soft-thresholding denoise; house prices in Beijing; support vector machine (SVM)

(责任编辑:汪 琴)