Hans Journal of Data Mining 数据挖掘, 2017, 7(2), 37-45
Published Online April 2017 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2017.72004

Hans汉斯

Analysis of Beijing Second-Hand House Price Based on Random Forest

Xiaotong Li, Xuan Guo, Chengjie Wang

College of Science, China University of Petroleum (Beijing), Beijing
Email: moonlane2009@126.com, 1540473460@qq.com, 1367593466@qq.com

Received: Apr. 5th, 2017; accepted: Apr. 27th, 2017; published: Apr. 30th, 2017

Abstract

With the development of economy and reducing of available land, the price of second-hand house is rising continuously. By the end of May 2016, average price of second-hand house in Beijing has been more than ¥60,000/m². Evaluating the price of second-hand house will not only produce important influence on residents' life, but also bring effective reference on the government's macroeconomic regulation and control. Current mathematical model about housing price includes linear regression model, neural network model (NN) and support vector machine model (SVM). In linear regression model, the suppose of linear relationship may cause more error. NN and SVM are proved to have poor explanatory. Based on the price of 16,795 second-hand houses in Beijing, the random forest model was established to study the influence factors of house price and the forecast of house price. Method of variance explanatory changes shows lat (Residential latitude), long (Residential longitude) and cate (Residential area) are the three main significant prediction variables on housing price, while random forest model picks up cate, lat and long to be the most important. Through analysis of OOB (out-of bag) samples, random forest gets a precision of 0.69 in second-hand housing forecast. Finally, put price data into NN and SVM model and forecast, precision 5.15 and 1.10 were got respectively. The result shows that random forest forecast is the best, followed by SVM. NN prediction does not apply to the second-hand house data in this paper.

Keywords

Second-Hand House, Housing Forecast, Bootstrap Sampling, Decision Trees, Random Forest Model

基于随机森林方法的北京市二手房价格研究

李晓童，郭　萱，王成杰

中国石油大学(北京)理学院，北京
Email: moonlane2009@126.com, 1540473460@qq.com, 1367593466@qq.com

收稿日期：2017年4月5日；录用日期：2017年4月27日；发布日期：2017年4月30日

文章引用: 李晓童, 郭萱, 王成杰. 基于随机森林方法的北京市二手房价格研究[J]. 数据挖掘, 2017, 7(2): 37-45.
https://doi.org/10.12677/hjdm.2017.72004

摘　要

随着经济的发展和可供开发土地的减少，二手房价一路飙高。截止到2016年5月底，北京城内六区二手房均价已超6万。对二手房价格进行评估预测将对居民生活产生重要影响，也可以给政府宏观调控提供一定参考。目前关于房价的数学模型多使用线性回归模型，神经网络模型和支持向量机模型。线性回归模型中对房价与预测变量线性关系的设定易造成较大误差，神经网络与支持向量机解释性较差。本文针对北京市16,795套在售二手房，对多类别变量建立随机森林模型，进行房价影响因素研究以及房价预测，通过方差解释性变化得到lat (小区所处纬度)，long (小区所处经度)和cate (小区所处区域)三个预测变量对房价的影响最为显著，通过随机森林变量重要性输出得到cate，lat和long对房价的影响最大。然后通过00B (out-of bag)样本得到随机森林二手房价格预测精度为0.69。最后将房价数据输入神经网络模型与支持向量机模型，得到房价预测精度分别为：5.15、1.10。结果表明，随机森林预测效果最佳；支持向量机模型次之，预测结果不够稳定；而神经网络预测误差较大，不适用于本文二手房价格预测。

关键词

二手房，房价预测，Boostrap抽样，决策树，随机森林

1. 引言

经济的快速发展使得城市可供利用开发的土地越来越少，房地产市场越来越成熟使得二手房市场交易日益活跃。今年三月北京市西城区文昌胡同的一处学区房卖出 46 万一平。有些房子是小朋友上好学校的通行证，可能是房子很贵的原因。同样房子周围有没有地铁也是影响房价的因素。通过对北京市二手房价格影响因素进行分析并对房价做出预测，为二手房价格评估提供理论依据和实践指导具有重要意义。对二手房的已有研究成果较多。仲小瑾[1] (2008)将多元线性回归模型应用于房价影响因素研究预测中。李菲等[2] (2004)使用灰色系统理论对线性回归模型进行完善和改进，建立房价预测模型。张辉[3] (2013)将非线性模型神经网络用于房价评估领域。陈静[4] (2008)将支持向量机模型用于西安市房价评估中。郭志强[5] (2013)将支持向量机用于房价预测中，发现预测结果明显优于岭回归与神经网络模型。这些研究成果中，线性回归模型遇到非线性问题会产生较大误差，神经网络稳定性较差以及对结果解释性较差，支持向量机解释性差。

随机森林由 Brieman 2001 年提出至今，已经被广泛应用于生态学、经济管理、生物医学、信用评价等领域。其不易出现过度拟合、很好的处理类别变量、解释性好、对噪声数据的容忍性、精度高等优点使其成为一种广泛使用的回归分类算法。本文用随机森林方法对北京市二手房价格影响因素进行分析并对房价做出预测，首先根据问题背景以及获取数据的局限性初步给出九个影响二手房价格因素，分别为：subway、school、long、lat、cate、bedrooms、halls、area、floor。基于 16,795 套在售二手房数据建立随机森林模型，由随机森林变量重要性输出以及逐步删除变量得到解释性变化值，从而得到影响北京市二手房价格的主要因素并进行分析，最后利用随机森林 OOB 样本数据对房价进行预测。同时本文还将神经网络模型、支持向量机模型作为对比模型，进行房价预测。发现在预测方面，随机森林有着更加精确的

预测效果。

2. 随机森林模型介绍

近年来，作为机器学习方法之一的随机森林受到越来越广泛的关注。随机森林[6]是一种统计学习理论，利用 bootstrap 抽样的方式从原始数据集中抽取多个样本，对每个 bootstrap 样本进行决策树建模，组合多个决策树投票得到最终预测结果。大量理论实践研究都表明随机森林具有很高的预测准确率，对异常值和噪声具有很好的容忍度，且不易出现过度拟合。随机森林作为一种非线性的建模工具，是目前数据挖掘、生物信息学最热门最前言的研究领域之一。

决策树是构成随机森林的基本单位，一个简单的决策树模型如图 1，其是一个树状结构，由根节点、中间节点以及叶结点组成。每一步根据变量的分类效果选择合适的划分，最终做出分类和预测

随机森林(图 2)是由决策树组成，通过组合多个决策树分类器进行分类和预测。其工作机制大致为：首先通过 bootstrap 抽样选择一系列训练集，在每个训练集上对特征进行随机的选取并通过基尼指标等其他指标对数据集进行合适的划分，生成一系列不剪枝的决策树。最后投票决定最优分类做出预测。

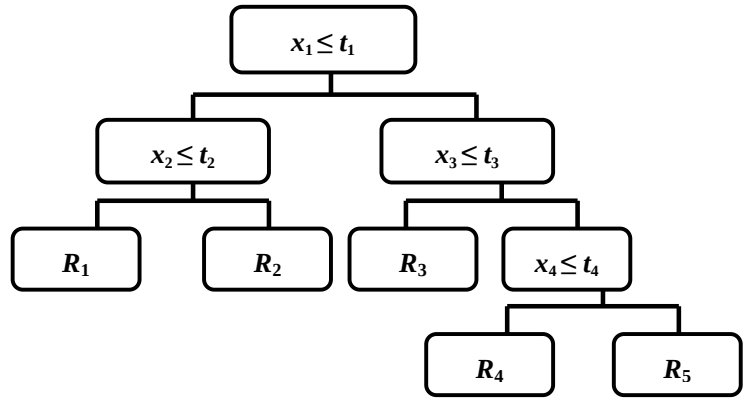

Figure 1. Decision model
图 1. 决策树模型

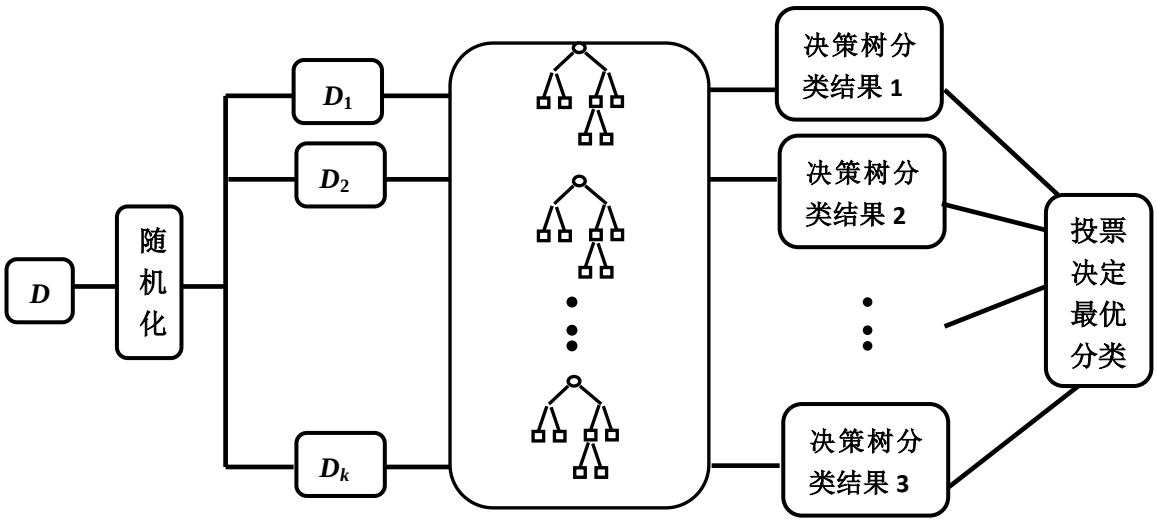

Figure 2. Random forest model
图 2. 随机森林模型

随机森林具体算法如下：

输入：1.训练集 $S=\{(x_i,y_i),i=1,2\cdots,n\},(X,Y)\in R^d\times R$

2.待测样本 $x_t\in R^d$

For: $i=1,2,\cdots,N_{tree}$

1) 对原式训练集 S Bootstrap 抽样，生成训练集 S_i

2) 使用 S_i 生成一棵不剪枝的树 h_i

a.从 d 个特征中随机选取 M_{try} 个特征

b.在每个节点上从 M_{try} 个特征依据 Gini 指标选取最优特征

c.分裂直到树生长到最大

End

输出：1.树的集合 $\{h_i,i=1,2,\cdots,N_{tree}\}$

2.对待测样本 x_t，决策树 h_i 输出 $h_i(x_t)$

回归：$f(x_t)=\frac{1}{N_{tree}}\sum_{i=1}^{N_{tree}}h_i(x_t)$

分类：$f(x_t)=majorityvote\{h_i(x_t)\}_{i=1}^{N_{tree}}$

由上述随机森林算法可得，随机森林的随机性主要体现在如下两方面：1) bootstrap 抽样产生的样本随机性。本文关于北京市 16,795 套二手房数据，通过 bootstrap 抽样，假设我们得到 500 个训练集，每个训练集中将近 37%的数据不会出现，训练集之间两两差异很大，由此对数据进行了充分利用；2) 在每个训练集上选择特征的随机性。在每个训练集上每一步进行特征选择时，不同于 bagging 的方法，随机森林会根据变量的个数确定选择几个特征。本文关于二手房数据的 9 个变量中，每一棵树生成时每一步划分我们选择 3 个变量，从 3 个变量中根据 Gini 指标确定最优的划分变量，生成不剪枝的决策树，依次生成一系列不剪枝决策树，相对于 bagging 方法，通过这样的特征选取进一步提高了数据的利用率，从而提高了预测精度。由这两点的随机性决定着随机森林的分类预测效果。

3. 二手房的随机森林模型

3.1. 变量的选取

本文主要对北京市二手房价格影响因素进行分析研究并对房价进行预测。收集某二手房中介网站 2016 年 5 月底北京城内六区(东城、西城、朝阳、海淀、丰台、石景山) 16,795 套在售二手房相关数据。

下面介绍本文各变量的选取：

1) 响应变量

本文选取在售二手房 Price (每平米的均价)作为响应变量。

2) 预测变量

美国学者 Butler 提出了影响房地产价格的三大特征变量[7]：区位特征，建筑特征以及邻里环境。区位特征指的是住宅小区位于城市哪个区域，包含与固定区位属性相关的一些特征。一般选取到城镇中心区的距离量化该特征。建筑特征简而言之即为住宅本身的客观状况，包括：户型、面积、建筑年龄、建筑结构、装修、车库等。邻里环境具体指住宅小区的人文环境、自然环境以及治安管理等。这三种特征变量包含着属性的隐含价格，因为消费者对于属性的支付意愿是从住宅价格间接得到的。

据此我们根据三大特征变量选择出九个变量作为本文的预测变量。区位特征选择：subway (是否地铁沿线)、school (是否为学区房)、long (所在小区所处的经度)、lat (所在小区所处的纬度)、cate (东城、西城、

海淀、朝阳、丰台石、景山)。建筑特征选择：bedrooms (卧室数)、halls (厅总数)、area (房屋总面积)、floor (basement, low, middle, high)。由于住宅小区及治安管理缺乏统计标准以及环境变量的数据获取途径有限。邻里环境变量没有出现在本文中。

3.2. 随机森林模型的建立

随机森林模型是 Breiman (2001)首次提出，通过建立一系列的决策树组成随机森林模型，最终投票做出最后的预测。该算法具有需要调整的参数较少、不必担心过度拟合、分类速度快、能高效处理大样本数据、能估计特征因素的重要性、很好的处理类别变量、有较强的抗噪声能力等优点。与线性回归相比，避免了线性回归事先假定的线性关系不符合实际造成较大误差的情况。且随机森林不用对函数形式事先进行假设，避免了假设误差。

运用随机森林方法进行二手房价格评估，随机森林可以处理分类和回归问题。本文对于二手房的研究属于回归预测问题。随机森林回归的基本思想是：首先利用自助抽样法，从原始数据集中抽取 B 个样本，且每个样本容量都与原始数据集相同；然后对 B 个样本分别建立 B 棵树，得到 B 个结果；最后，对这 B 个结果取平均值得到最终的预测结果。基于随机森林的二手房价格评估模型计算如下：

二手房的随机森林模型由 B 棵树组成，$\{F_1(X), F_2(X), \cdots, F_B(X)\}$，其中 $X = \{x_1, x_2, \cdots, x_p\}$ 是二手房的 P 维特征向量。结果会产生 B 个结果 $\hat{Y}_1 = F_1(X), \hat{Y}_2 = F_2(X), \cdots, \hat{Y}_B = F_B(X)$。其中，$\hat{Y}_b, (b = 1,2,\cdots,B)$ 是第 b 棵树的预测结果。对于回归问题预测值为所有树预测结果的平均。算法流程如下：

1) 原始数据含样本量为 16,795，应用 bootstrap 方式抽样选择 500 个样本集，构建 500 棵决策树。每次抽样未被抽到的样本构成 OOB 样本作为随机森林的验证样本。

2) 样本中变量个数为 9，每一棵决策树每一个节点随机选择 m_{try} 个变量进行基尼指标计算，确定合适的变量得到合适的划分。使用随机森林做回归时，通常选取 $m_{try} = P/3$。本文每一次划分选择 3 个变量。

3) 每一棵决策树生长到最大，无需进行剪枝，重复上述步骤直到生成 500 棵决策树。

通过如上步骤，建立得到二手房的随机森林价格评估模型，将 OOB 样本输入随机森林模型得到房价预测精度。

3.3. 特征变量重要性评价

随机森林可以给出变量重要性排序，本文据此得出影响二手房价格的重要预测变量。其次，本文通过依此删除预测变量的方式计算方差解释性差值，得到变量重要性排序。删除某个变量后解释性差值变化越大，证明这个变量越重要；解释性差值变化越小，证明这个变量越不重要。

记删除变量后方差的解释性为：$w_j \cdots (j = 1,2,\cdots,9)$

方差解释性变化为：$t_i \cdots (i = 1,2,\cdots,9)$

变量分别如表 1。

为了提高计算准确性，随机森林运行十次得到方差解释性如表 2。

方差平均解释性为：85.12%。

逐个删除自变量，输入随机森林模型，方差解释性如表 3。

方差的解释性变化如表 4。

由此可得，可以按照重要性将变量分为三个层次：第一层次包括 lat、long、cate 三个方差的解释性差值最大的变量，这表明大多数人选择二手房时首先考虑房子所在的地理位置(纬度、经度和区域)。选择了房子的地理位置后，第二层次变量包括房子的总面积(area)以及卧室的数目(bedrooms)。第三层次的学区房和是否临近地铁这两个变量方差解释性差值较小，为重要性相对较弱的变量，分析其原因，是否临

Table 1. Explaining variable
表 1. 预测变量

x1	Subway	x2	school
x3	Long	x4	lat
x5	Cate	x6	bedrooms
x7	Halls	x8	area
x9	Floor		

Table 2. Variance of explanatory
表 2. 方差解释性

次数	解释性	次数	解释性
1	85.10%	6	84.90%
2	85.22%	7	84.85%
3	85.30%	8	85.26%
4	84.73%	9	85.65%
5	85.37%	10	84.80%

Table 3. Erase variables of variance explanatory
表 3. 逐个删除变量方差解释性

w1	82.95%	w2	81.92%
w3	74.91%	w4	71.95%
w5	76.16%	w6	80.92%
w7	81.72%	w8	79.97%
w9	81.19%		

Table 4. Changes of variance explanatory
表 4. 方差解释性变化

t1	2.17%	t2	3.20%
t3	10.21%	t4	13.17%
t5	8.96%	t6	4.20%
t7	3.40%	t8	5.15%
t9	3.93%		

近地铁是房子所处地理位置的一部分因素，大多数情况下可以由房子的地理位置确定，因此作为单独变量影响较小；而学区房受众群体比较单一，其重要性只针对有孩子需要上学的家庭，样本较大时这种重要性会被减弱。

随机森林输出的变量重要性如表 5。

由表 5 可得，cate，lat 和 long 同上述方差解释性差值一样，为最重要的三个变量，表明大多数人选择二手房时首先考虑房子所在的地理位置。school，subway 和 area 为接下来重要的变量。这个结果与上述方差解释性得到的结果具有大致相同的趋势。

Table 5. Output of RF variable importance
表 5. RF 变量重要性输出

变量	节点纯度
CATE	1.77E+12
Bedrooms	9.65E+10
Halls	8.75E+10
AREA	4.30E+11
floor	1.09E+11
Subway	1.09E+11
School	6.59E+11
LONG	8.73E+11
LAT	1.06E+12

3.4. 二手房房价预测

通过 bootstrap 抽样，未被抽到的样本组成了 B 个袋外数据(out-of-bag, OOB)，构成 OOB 样本。每次 bootstrap 抽样，将近 37%的样本不会被抽中。本文将入袋样本作为测试集，将袋外样本作为验证集。采用下述的方式衡量房价的预测精度：

$$ESS = \sum_{i=1}^{n}\left(P_{wyi} - P_{wsi}\right)^2 \tag{1}$$

$$J = \sqrt{ESS} \tag{2}$$

其中 n 为 16,795 套二手房数据的袋外数据量。 P_{wyi} 为袋外数据的预测价格， P_{wsi} 为袋外数据的实际价格。 ESS 为残差平方和。J 为残差平方和取平方根。

随机森林每一次 bootstrap 抽样，会产生不同的 OOB 样本，不同的 OOB 样本计算 ESS 会得到不同的预测精度，为了保证预测准确性，对十次 bootstrap 得到的袋外数据计算预测误差并取平均，为了方便与下文其他模型对比，我们取预测误差平均的10^{-6}，计算结果如表 6。

为了更加直观的看到随机森林的预测效果，我们使用 R 软件在 16795 个数据集的 OOB 样本中随机抽取 15 个样本，得到其预测价格与实际价格并计算预测误差如表 7。我们看到预测误差基本可以控制在 10%左右，说明随机森林预测效果良好。

4. 模型对比

分类和回归模型使用较好且常用的有神经网络模型与支持向量机模型。本文将数据输入这两个对比模型得到预测误差如下：

计算得到支持向量机十次的预测误差并取平均如表 8。

计算得到神经网络十次的预测误差并取平均如表 9。

本文关于北京市二手房数据我们得到方差的解释性达到 85.12%，表明所得数据里包含着大量可提取的有效信息，进一步变量重要性的输出对预测精度高做出合理的解释。将北京市二手房的 9 个预测变量分为三个层次，第一层次包括 lat、long 和 cate，三个表明房屋地理位置的变量；第二层次包括 area 和 bedrooms，两个表明房屋建筑特征的变量；第三层次包含 subway、school 等重要性相对较弱的变量。相对于神经网络模型和支持向量机模型直接给出预测精度，随机森林变量重要性的输出对房价进行了合理

Table 6. Prediction accuracy of RF model OOB sample
表 6. RF 模型 OOB 样本预测精度

次数	J	次数	J
1	697303	6	646405.7
2	700391.7	7	665065
3	711579.9	8	707365.6
4	681810	9	680221.9
5	698288	10	690601.6
平均	687903.2		0.687903

Table 7. Some predictions of house prices
表 7. 房价部分预测情况

序号	实际价格	预测价格	预测误差	序号	实际价格	预测价格	预测误差
10736	56389	63590.84	12.77%	5735	42639	39447.53	7.48%
6690	45376	39248.68	13.50%	13378	50477	49622.64	1.69%
365	66679	62198.62	6.72%	7204	38637	38118.09	1.34%
1876	41243	43597.08	5.71%	9673	58352	59597.04	2.13%
9823	92527	99416.12	7.45%	11472	86402	78716.34	8.90%
8017	75377	74278.09	1.46%	11522	70198	56844.28	19.02%
4030	36539	34855.84	4.61%	14333	114183	88392.41	22.59%
826	39797	38470.58	3.33%				

Table 8. Prediction accuracy of SVM model
表 8. SVM 模型预测精度

次数	J	次数	J
1	1101259	6	1113980
2	1093232	7	1089112
3	1111184	8	1119870
4	1120403	9	1104404
5	1089643	10	1096588
平均	1103968		1.103968

Table 9. Prediction accuracy of NN model
表 9. NN 模型预测精度

次数	J	次数	J
1	5166610	6	5149572
2	5160513	7	5130260
3	5182932	8	5196509
4	5165548	9	5125228
5	5113012	10	5137804
平均	5152799		5.152799

的解释。由上神经网络和支持向量机对比模型可得，支持向量机模型预测的误差仅次于随机森林模型预测误差，但误差较大，约为随机森林误差的一倍。神经网络的误差较大，不适合于本文二手房房价评估模型。

5. 总结

本文构建了二手房价格评估的随机森林模型。在三大特征变量中选择了 9 个预测变量，对北京市城内六区 16,795 套在售的二手房数据进行了房价影响因素以及房价预测研究。研究表明，cate，lat 和 long 为影响房价的最重要变量。进一步本文利用 OOB 样本实现了对随机森林模型预测精度的外推，得到了随机森林有着较好的预测精度。最后本文引入对比模型神经网络模型以及支持向量机模型对房价进行预测，得到支持向量机模型的预测效果仅次于随机森林模型的预测效果，而神经网络模型预测误差较大，不适用于本文的房价预测。

基金项目

中国石油大学(北京)本科教育教学改革项目，项目编号 21G16091。

参考文献 (References)

[1] 仲小瑾. 基于多元线性回归分析法的房地产价格评估[J]. 商业时代, 2014: 133-134.

[2] 李菲, 孙文彬. 灰色理论在商品住宅价格预测中的应用[J]. 辽宁工程大学学报, 2004, 6(3): 271-273.

[3] 张辉. 关于多当今社会 BP 神经网络的房地产价格评估与研究方向[J]. 房地产导刊, 2013.

[4] 陈静. 基于支持向量机的房地产估价方法研究[D]: [硕士学位论文]. 西安: 长安大学, 2008.

[5] 郭志强. 基于支持向量机回归的房地产批量估价[D]: [硕士学位论文]. 广州: 暨南大学, 2013.

[6] James, G. (2014) An Introduction to Statistical Learning with Applications in R. University of Southern California, 303-324.

[7] 杨沐晞. 基于随机森林模型的二手房价格评估研究[D]: [硕士学位论文]. 长沙: 中南大学, 2012.