Advances in Applied Mathematics 应用数学进展, 2021, 10(9), 3134-3140
Published Online September 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.109327

Hans汉斯

基于随机森林的猪肉价格预测模型

黄　琪，董帮强

东华理工大学理学院，江西　南昌
Email: 852410021@qq.com, 407654553@qq.com

收稿日期：2021年8月17日；录用日期：2021年9月9日；发布日期：2021年9月22日

摘　要

中国是世界上最大的猪肉生产国和消费国，也是生猪养殖规模最大的国家。在中国的所有肉类消费中，猪肉消费一直领先其他肉制品，近年来生猪价格波动频繁，会对整个养猪业乃至社会造成巨大影响。因此探究一种可以准确预测猪肉价格的模型对生猪市场的研究和生产都具有重要意义。随机森林是以K个决策树为基本分类器，进行集成学习后得到的组合分类器，可以解决数据多模态问题。考虑到猪肉价格与其他因素之间的复杂多模态非线性关系，故本文使用随机森林对猪肉价格进行预测。针对收集的猪肉价格影响因素(如玉米价格，牛肉价格等)，建立多棵决策树构建随机森林模型，对猪肉价格实现精准预测。同时进行了对比实验，对比决策树、支持向量机预测模型，实验结果表明基于随机森林的预测价格数据和真实价格数据拟合效果最好。

关键词

猪肉价格预测，随机森林，支持向量机，决策树

Pork Price Prediction Model Based on Random Forest

Qi Huang, Bangqiang Dong

School of Science, East China University of Technology, Nanchang Jiangxi
Email: 852410021@qq.com, 407654553@qq.com

Received: Aug. 17th, 2021; accepted: Sep. 9th, 2021; published: Sep. 22nd, 2021

Abstract

China is the world's largest producer and consumer of pork, and it is also the country with the largest scale of pig farming. Among all meat consumption in China, pork consumption has always

文章引用: 黄琪, 董帮强. 基于随机森林的猪肉价格预测模型[J]. 应用数学进展, 2021, 10(9): 3134-3140.
DOI: 10.12677/aam.2021.109327

been ahead of other meat products. In recent years, the price of live pigs has fluctuated frequently, which will have a huge impact on the entire pig industry and society. Therefore, exploring a model that can accurately predict the price of pork is of great significance to the research and production of the live pig market. Random forest is based on K decision trees as the basic classifier, and the combined classifier is obtained after ensemble learning, which can solve the data multi-modal problem. Considering the complex multi-modal nonlinear relationship between pork price and other factors, this paper uses random forest to predict pork price. According to the collected pork price influencing factors (such as corn price, beef price, etc.), establish multiple decision trees to construct a random forest model to realize accurate prediction of pork price. At the same time, a comparative experiment was carried out to compare the prediction model of decision tree and support vector machine. The experimental results showed that the predicted price data based on random forest and the real price data have the best fit.

Keywords

Pork Price Prediction, Random Forest, Support Vector Machine, Decision Tree

1. 引言

中国是世界上最大的猪肉生产国和消费国，据国家统计局数据，2020 年中国生猪出栏量高达 5.2 亿头，占世界总产量的 52%，2020 年生猪产业值占农业总产值的比重达 18% [1]。猪肉也在畜牧产品中占有主要地位，我国猪肉的消费量占世界猪肉消费量的一半，因此其价格的变化影响着我国居民的生活质量水平。所谓“粮猪安天下”，足见养猪业作为中国食品安全的基础产业和战略产业，不仅关系到国民经济发展，也与社会稳定息息相关。近年来，猪肉的价格浮动区间较大，且猪肉价格受到饲养成本、市场需求、疾病等诸多不确定因素的影响而出现难以把控和预测的特点，这不仅使养猪户出现不敢养和积极性不高的问题同时也给其他相关产业带来了一定影响。因此，预测与稳定猪肉市场的价格会对整个养猪行业健康发展有重要的意义[1]。

社会科学研究者一般通过市场调研的方法进行定性的分析，没有进行定量分析，结果不准确。猪肉价格预测本质上是一个回归问题，即通过易测得的变量来预测不易获取的变量。随着数据驱动、人工智能的发展，大数据分析的方法在预测猪肉价格、预测产量等方面有了深入研究。多元统计回归就是其中一个分支。最小二乘回归[2]、岭回归[3]、主元回归[4]、偏最小二乘回归[5]等多元统计回归模型都有应用于价格分析、预测中，但它们能进行预测的前提是数据之间是线性关系，而猪肉价格预测是一个复杂的多维非线性问题[6] [7]，因此线性方法无法挖掘变量间的非线性特性。

神经网络可以拟合任意非线性函数，能深入挖掘变量间的非线性特性。李阳等人构建了基于灰色预测－反向传播(GM-BP)神经网络预测模型对猪肉价格进行了有效预测[8]。刘青松等人提出基于自回归－循环神经网络(AR-RNN)的多变量水位预测模型[9]。赵圆芳等人提出基于 LSTM 循环神经网络的质量预测方法[10]。以上文献表明用神经网络做预测可以提高预测结果的准确性，但神经网络模型是一个黑箱模型，不具备可解释性。

机器学习不仅能解决非线性模型的问题，相比神经网络具有更好的解释性，而且机器学习不需要大量数据也能预测拟合地很好，因此，机器学习在价格回归等方面有着比较广泛的应用。张方怡等人以贝

叶斯网络为理论依据，利用 Hugin Lite 软件拟合猪肉合格率模型[11]。姜百臣等人基于集成经验模态分解的遗传算法改进支持向量机来预测猪肉价格[12]。Jumin Ellysia 等人根据马来西亚收集的数据，应用决策树回归来预测太阳辐射的变化[13]。陈帅通过研究相关文献发现生猪价格受各种因素影响，运用逐步回归、随机森林及神经网络等方法综合分析得出影响生猪价格的几个主要因素[14]。

考虑到猪肉价格与其他因素之间的复杂多模态非线性关系，本文使用随机森林对猪肉价格进行预测。针对收集的猪肉价格影响因素(如玉米价格，牛肉价格等)，建立多棵决策树构建随机森林模型，对猪肉价格实现精准预测。

2. 猪肉价格影响因素分析

猪肉价格市场是开放和多元性的系统，经过阅读大量研究文献与市场调查实证分析可知影响猪肉价格的因素总体上可分为以下四个方面：供给关系、需求关系、成本、市场外因[15]。在供给方面主要受到猪肉产量、生猪存栏量和出栏量等。需求方面主要受到居民消费水平、其他肉食产品价格等影响。成本方面主要有仔猪费用、育肥猪饲料、人工成本、生猪价格、玉米等影响。市场外因主要包括自然灾害原因、疫病等影响。影响猪肉价格波动的原因是多维和非线性的，在此本文结合前人研究成果[16]，特选取生猪价格、仔猪价格、猪粮比价格、豆粕价格、玉米价格、鸡肉价格、牛肉价格、羊肉价格、育肥猪饲料价格、鸡蛋价格、活鸡价格，共 11 种影响因子进行猪肉价格预测研究。

3. 方法介绍与数据来源

猪肉价格与其影响因素之间是复杂且非线性的关系，随机森林又具有较强的高维数据处理能力，还能对各个影响因素的重要性进行解释，适用于猪肉价格的预测与研究。故本文采用随机森林进行猪肉价格预测。

3.1. 决策树

决策树作为随机森林算法的基学习器，是研究随机森林算法必不可少的部分。随机森林就是由多棵决策树构成的。决策树中分别有根节点、叶节点和中间节点。决策树在构建时，由根节点开始分裂，分裂经过多个中间节点，最终到达叶节点。在这个过程中，决策树中的节点即为各个特征，从各节点分裂而出的路径表示特征可能选取的值。决策树的输出规则是唯一的，即最终输出值是唯一的。换句话说，从根节点开始仅能到达唯一的叶子节点，因此可以用来分类及预测[17]。

在决策树的分裂过程中，属性分裂方式是择优录取即选择分裂时结果最好的属性进行分裂。比较分裂结果的好坏有多种方法，依据这些方法可以分为 CLS、ID3、C4.5、CART 等节点分裂算法[17]。本文主要使用的是 ID3 节点分裂算法来进行建模预测。

ID3 算法结合了信息熵理论，并将其属性节点分裂的方法。它的主要规则是：首先计算各个属性的信息增益，选择信息增益大的节点进行分裂。设数据集为 D，包括了 n 个不同的类别 C_i，其中 $C_{i,D}$ 是数据集 D 中 C_i 类元组的集合，$|D|$ 是 D 元组的个数，$|C_{i,D}|$ 是 $C_{i,D}$ 元组的个数。

D 中元组的期望信息可以表示为

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{1}$$

其中，p_i 为 C_i 类元组出现的频率。

若用属性 A 将 D 分为若干个子集 $\{D_1, D_2, \cdots, D_n\}$，基于 A 划分 D 的期望信息表示为：

$$Info_A(D) = \sum_{j=1}^{m} \frac{|D_i|}{|D|} * Info(D_j) \tag{2}$$

其中，$\frac{|D_i|}{|D|}$充当第 j 个划分的权值。$Info_A(D)$越小，划分的纯度越高。

根据上两式可得到信息增益为：

$$Gain(D) = Info(D) - Info_A(D) \tag{3}$$

3.2. 随机森林算法原理

随机森林具备两种随机思想，分别是样本选取时的 bagging 思想和特征选择时的随机子空间思想。对每一个 bootstrap 样本构建决策树，然后将所有的决策树的预测结果集成进而生成随机森林的预测结果。在集成决策树结果时，回归算法就对决策树的预测结果取平均数，分类算法就采取众数投票的方式。

随机森林是由多棵决策树$\{h(x,\theta_t), t=1,2,\cdots,T\}$组合形成的模型。其中 θ_t 为服从独立分布的随机变量，x 为自变量，T 为决策子树的个数。

(1) 分类模型预测结果为：

$$\bar{H}(x) = \arg\max \sum_{i=t}^{T} I\left(h(x,\theta_t) = Y\right) \tag{4}$$

上式中，$\bar{H}(x)$表示分类结果，Y 表示分类类型，I 表示示性函数。

(2) 回归模型预测结果为：

$$\bar{h}(x) = \frac{1}{T}\sum_{i=t}^{t}\{h(x,\theta_t)\} \tag{5}$$

上式中，$\bar{h}(x)$表示回归预测结果，$h(x,\theta_t)$表示基于 x 和 θ 的输出。

3.3. 随机森林构建过程

随机森林构建步骤为：

(1) 利用 bagging 思想进行有放回抽样，要产生 N 棵决策树，就需要从原始数据集合中有放回的抽取 N 次形成 N 个样本子集，这其中每个样本子集所包含的样本量大概是原始数据集合样本量的 2/3 [17]。

(2) 对所抽取的训练子集，利用随机子空间思想，随机抽取 f 个特征作为特征子空间，从特征子空间中挑选最优特征并以此开始进行节点分裂，进而构建决策树。在节点分裂时，对于回归模型，则基于均方误差建立回归树；对于分类模型，则基于基尼指数建立分类树[17]。

(3) 重复步骤(1)、(2)，生成 T 棵决策树。对每一颗决策树，任由其生长，不对其进行剪枝，最终由这 T 棵决策树构成整个随机森林[17]。

(4) 综合 T 棵决策树的预测结果，汇总得到随机森林的预测结果。对于回归模型，采取取平均方式；对于分类模型，采取投票方式[17]。

3.4. 数据来源

为了有效保证本次预测结果的准确性，在选取猪肉价格等相关数据因素时，数据均来源于中国畜牧业信息网、布鲁克农业数据平台并结合实地调研与文献统计而来。本文数据选取 2010 年 1 月~2020 年 12 月生猪价格、仔猪价格、猪粮比价格、豆粕价格、玉米价格、鸡肉价格、活鸡价格、牛肉价格、羊肉价格、育肥猪饲料价格、鸡蛋价格，这些月度价格为研究对象，共计 132 组数据。

4. 预测结果及分析

本文选取 2010 年 1 月~2020 年 12 月生猪价格、玉米价格等 11 类月度价格为研究对象，随机挑选 60%

组数据作为训练样本，剩下 40%组数据作为测试样本，利用随机森林模型对其进行预测、检验。

4.1. 模型评估指标

为了评价预测性能，本文使用均方误差(MSE)作为模型的评价准则，定义如下：

$$\mathrm{MSE}=\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2 \tag{6}$$

其中：y_i 为真实值，$\hat{y}_i$ 为预测值，m 为数据长度。显然，均方误差值越小，预测精度越高，预测模型效果越好。

4.2. 基于随机森林的猪肉价格预测建模

将 132 组数据随机打乱，其中随机挑选 60%组数据用作训练样本，剩下的 40%组数据用作测试样本，对猪肉价格预测模型进行检验。这里设置随机森林的决策树模型数目为 100，即将 100 棵决策树的效果进行集成，将不同决策树的结果整合以获得最终随机森林的结果。此外，我们建立了单棵决策树模型和支持向量机模型对猪肉价格进行预测，并于随机森林的效果进行比较。

4.3. 模型的效果与分析评价

利用 python 平台分别编写关于决策树算法、支持向量机算法、随机森林算法三种预测方法的相关程序代码，在运行计算后并将其预测结果绘制如图 1 所示。通过对输出的预测猪肉价格与真实的猪肉价格计算得到三种预测模型的均方误差如表 1 所示。

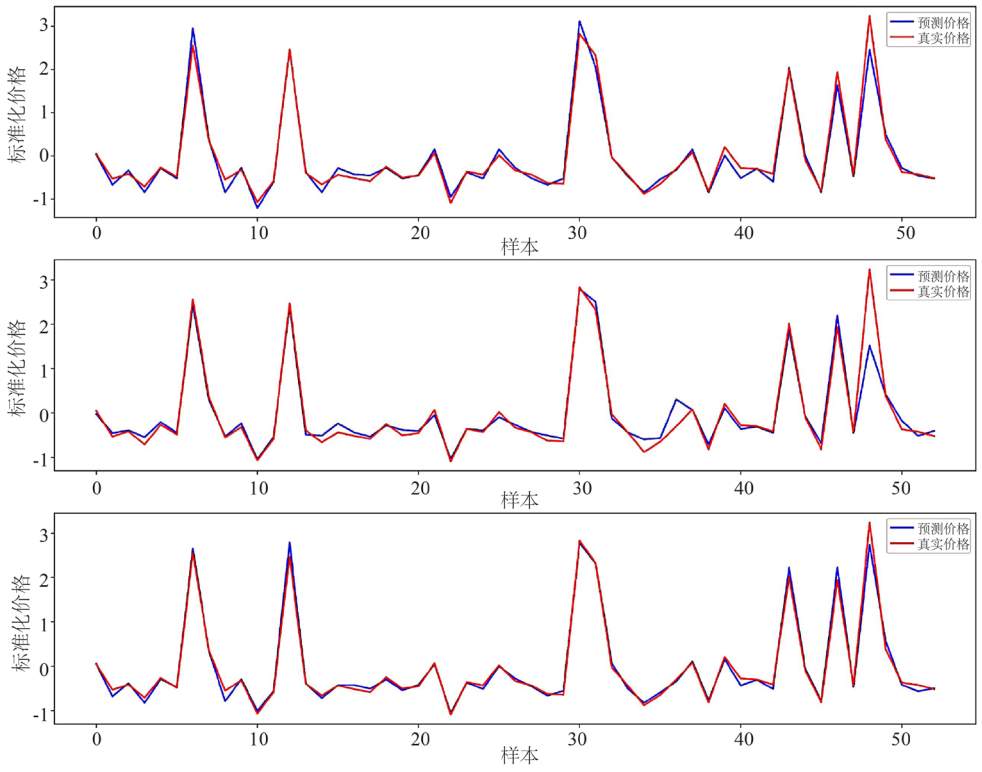

Figure 1. Comparison chart of three model predictions: decision tree, support vector machine, and random forest

图 1. 三种模型预测对比图：决策树、支持向量机、随机森林

Table 1. The mean square error of the three forecasting models
表 1. 三种预测模型的均方差

方法	均方误差
决策树	0.097
支持向量机	0.123
随机森林	0.081

从图 1 中可以看出，在验证的 40%猪肉数据中，随机森林的猪肉价格预测值和真实值是最接近的，而决策树预测在[10, 15]，[45, 50]这两个区间偏差较大，支持向量机预测在[46, 50]预测偏差较大。由此可以得出随机森林在猪肉价格预测建模中做的最好。而从表 1 可以看出，随机森林预测模型的均方误差是 0.081，在三种模型中，其均方误差是最小的，也即真实值和预测值差距最小。

5. 结论

由于猪肉价格与其他因素之间的复杂多模态非线性关系，本文采取了随机森林预测模型对猪肉价格进行精准预测，并同时使用了决策树模型、支持向量机模型进行对比实验，实验结果表明，使用随机森林预测猪肉价格，其真实价格和预测价格拟合程度最优，且均方误差最小，达到了预测精度的要求范围，因此可以说明该模型对于我国猪肉价格的预测具有较大的适用性。

对于猪肉价格的影响不仅包括供给关系、需求关系、成本、市场外因的影响，同时也容易受到疫病的影响。因此在实际的猪肉价格预测过程中还存在一些“理想处理因素”和主观性因素。所以在以后的研究过程中应当提高数据的准确性以及处理方法的融合性，来进一步提高预测的准确度，提高猪肉价格预测的准确率。

参考文献

[1] 马再兴. 基于动态贝叶斯网络的生猪价格与产量预测[D]: [硕士学位论文]. 武汉: 华中农业大学, 2019.

[2] 石雪涛, 朱帮助. 基于相空间重构和最小二乘支持向量回归模型参数同步优化的碳市场价格预测[J]. 系统科学与数学, 2017, 37(2): 562-572.

[3] 陈海鹏, 卢旭旺, 申铉京, 杨英卓. 基于多元线性回归的螺纹钢价格分析及预测模型[J]. 计算机科学, 2017, 44(11): 61-64.

[4] 董京铭, 刘瑞翔, 马晨晨, 等. 利用主成分回归方法预估连云港地区水稻气象产量[J]. 江苏农业学报, 2021, 37(3): 606-612.

[5] 苗田恬, 刘思琦. 基于偏最小二乘回归的农产品种植评价研究[J]. 商场现代化, 2019(15): 19-20.

[6] 张宇青, 周应恒, 易中懿. 中国生猪出栏价格波动的非线性特征分析与预测[J]. 统计与决策, 2015(1): 141-143.

[7] 张莹莹. 基于 ARIMA 模型的中国猪肉价格预测[J]. 商展经济, 2020(15): 21-23.

[8] 李阳, 王晓光. 基于 PCA-GM-BP 神经网络的猪肉价格预测分析[J]. 数学的实践与认识, 2021, 51(5): 56-63.

[9] 刘青松, 严华, 卢文龙. 基于 AR-RNN 的多变量水位预测模型研究[J]. 人民长江, 2020, 51(10): 94-99.

[10] 赵圆芳, 高媛, 钱峰, 亢欣欣. 应用 LSTM 网络的缸体压铸质量预测[J]. 机械设计与制造, 2021(7): 229-232.

[11] 张方怡, 董庆利, 等. 基于贝叶斯网络的猪肉合格率的模型构建[J]. 食品工业科技, 2012, 33(10): 52-54+93.

[12] 姜百臣, 冯凯杰, 彭思喜. 基于改进持支向量机的猪肉价格预测研究[J]. 广东农业科学, 2018, 45(12): 158-164.

[13] Jumin, E., Basaruddin, F.B., *et al.* (2021) Solar Radiation Prediction Using Boosted Decision Tree Regression Model: A Case Study in Malaysia. *Environment Science and Pollution Research*, **28**, 26571-26583. https://doi.org/10.1007/s11356-021-12435-6

[14] 陈帅. 我国生猪价格波动影响因素分析[J]. 农村经济与科技, 2019, 30(21): 63-65.

[15] 李苏, 宝哲, 多春梅. 局部均衡理论视角的中国猪肉产品供需预测[J]. 黑龙江畜牧兽医, 2019(4): 1-7.

[16] 任青山, 方逵, 朱幸辉. 基于多元回归的 BP 神经网络生猪价格预测模型[J]. 江苏农业科学, 2019, 47(14): 277-281.

[17] 徐艳平. 基于改进的随机森林算法的城市空气质量预测模型[D]: [硕士学位论文]. 重庆: 重庆工商大学, 2021.