

doi: 10.12133/j.smartag.2020.2.3.202008-SA003

基于Lasso回归和BP神经网络的蔬菜短期价格 预测组合模型研究

喻洸舸^{1,2}, 吴华瑞^{1,2,3*}, 彭程^{1,2,3}

(1. 北京市农林科学院, 北京 100097; 2. 国家农业信息化工程技术研究中心, 北京 100097;

3. 农业农村部农业信息技术重点实验室, 北京 100097)

摘要: 蔬菜是居民生活饮食的重要组成部分, 蔬菜价格预测存在着价格波动幅度大、影响因素复杂多样、精度不高等难点。本研究以黄瓜为研究对象, 分析了影响黄瓜价格的供给、需求、流通等因素, 引入Lasso回归模型对影响因素进行筛选, 获得12项关联度较大的因素。在此基础上, 构建了一种基于影响因素的Lasso回归方法与BP神经网络相结合的组合模型(L-BPNN), 开展黄瓜短期价格预测, 并与Lasso回归模型、BP神经网络模型、RBF神经网络模型等回归分析和智能分析方法等进行了对比验证研究。结果表明: 使用L-BPNN模型预测黄瓜价格, 其平均相对误差最小, 仅为0.66%, 比Lasso回归模型、BP神经网络模型和RBF神经网络模型分别低64.52%、82.11%和86.2%, 具有较高的预测精度。本研究结果实现了黄瓜的短期价格预测, 也可推广到其他蔬菜品种, 对于保障菜农收入、稳定蔬菜市场价格等具有重要意义。

关键词: 蔬菜; 影响因素; 价格预测; 组合模型; Lasso回归方法; BP神经网络; RBF神经网络

中图分类号: S-126

文献标志码: A

文章编号: 202008-SA003



引文格式: 喻洸舸, 吴华瑞, 彭程. 基于Lasso回归和BP神经网络的蔬菜短期价格预测组合模型研究[J]. 智慧农业(中英文), 2020, 2 (3): 108-117.

YU Weige, WU Huarui, PENG Cheng. Short-Term Price Forecast of Vegetables Based on Combination Model of Lasso Regression Method and BP Neural Network[J]. Smart Agriculture, 2020, 2 (3): 108-117. (in Chinese with English abstract)

1 引言

蔬菜是农业产业支柱, 也是国民生活的必需品, 其价格影响着国民生活水平。蔬菜价格容易受到多种因素影响, 近几年部分品种的蔬菜价格波动幅度较大, 波动较为频繁。蔬菜价格的异常波动一方面对菜农的经济利益造成了损失, 另一方面也影响了居民日常饮食与生活质量, 特别是

低收入人群的生活品质。因此, 开展蔬菜价格预测, 进而知道农业生产具有重要意义。

当前, 国内外学者通过数理统计预测法、计量经济预测法、智能分析法、组合模型法等对蔬菜价格预测展开探索, 取得了有效进展^[1-3]。如Ahinli^[4]应用霍尔特-温特斯(Holt-Winters)和整合移动平均自回归模型(Autoregressive Integrated Moving Average Model, ARIMA)方法成

收稿日期: 2020-08-13 修订日期: 2020-09-23

基金项目: 国家大宗蔬菜产业技术体系岗位专家项目(CARS-23-C06)

作者简介: 喻洸舸(1993—), 男, 硕士, 研究方向为农产品市场价格分析。E-mail: yuwg@nercita.org.cn。

*通讯作者: 吴华瑞(1975—), 男, 博士, 研究员, 研究方向为农业大数据与智能系统。电话: 010-51503921。E-mail: wuhr@nercita.org.cn。

功预测马铃薯消费价格。Ghosh 等^[5] 和 Paredes-garcia 等^[6] 分别使用带有输入序列的一般 ARIMA 模型和季节性自回归整合移动平均模型实现对蔬菜价格进行预测。这些主要基于时间序列方法对农产品价格进行预测, 越来越多的学者运用智能分析方法对农产品价格预测进行探索。陈芳和楼文高^[7] 探究了蔬菜日价格数据的预测, 通过搭建广义回归神经网络模型 (Generalized Regression Neural Networks, GRNN), 获得了较小的预测误差。叶露等^[8] 分别使用粒子群改进的 BP 神经网络、RBF 神经网络建模进行蔬菜价格预测, 发现两者各有优势, 两种方法线性结合后会有更优的表现。刘合兵等^[9] 基于小波变换和 BP 神经网络构建农产品价格组合预测模型, 实现菠菜、大白菜等 5 种蔬菜月度价格预测。贾宁和郑纯军^[10] 设计了一种双重注意力机制与长短期记忆网络融合的神经网络模型, 通过模拟可以较高精度地预测一周的蔬菜价格整体走势。Xiong 等^[11] 提出了一种基于季节趋势和极端学习机器的混合方法 (Seasonal-Trend Loess and Extreme Learning Machines, STL-ELM), 用于季节性蔬菜价格的短期、中期和长期预测。Zhang 等^[12] 引入分位数回归-径向基函数神经网络模型 (Quantile Regression-radial Basis Function, QR-RBF) 对中国国内大豆月价格数据进行了分析。

以上国内外学者的研究使用传统的计量经济方法和数理统计方法, 或借助智能预测手段, 开展了许多有效的蔬菜价格预测探索, 但是预测方法大多数局限于基于时间序列的价格预测, 而忽视了农产品价格影响因素, 基于影响因素的蔬菜价格预测的研究还比较少。本研究以黄瓜为研究对象, 总结分析黄瓜价格各类影响因素, 引入 Lasso 回归模型对影响因素进行筛选, 获得关联性较大的黄瓜价格影响因素, 在此基础上, 引入 BP 神经网络, 构建了 Lasso 回归方法与 BP 神经网络相结合的组合模型 (L-BPNN), 对黄瓜价格进行短期预测 (一般预测期在 1 年以下的预测为短期预测^[13]), 通过多种模型对比证明了该模型

具备预测精度高、预测效果稳定的优势。

2 蔬菜价格影响因素分析

2.1 蔬菜价格影响因素

与一般商品相比, 蔬菜具有季节性、鲜活性等特征, 蔬菜价格的影响因素主要包括以下四个方面^[14-16]。

(1) 供给因素。它是蔬菜价格决定中的基础部分, 在蔬菜的市场交易中, 生产者的产出直接影响到市场的定价。从蔬菜生产过程来看, 播种面积和亩产量是两个重要因素, 在市场需求未明显变化时, 蔬菜产量下降通常带来价格上涨, 而蔬菜产量过剩时, 往往会带来价格的下跌。从经济运行原理来看, 蔬菜生产过程中投入的各项成本也会直接影响到最终定价。蔬菜产业成本包括土地成本、生产资料成本和用工成本等。

(2) 需求因素。从市场需求来看, 蔬菜产业涉及到消费者的消费水平以及对蔬菜的消费需求。通过文献查阅和综合分析, 蔬菜消费水平可以从以下因素考虑: 城镇居民人均可支配收入、城镇居民人口数量、城镇蔬菜消费价格指数等。另外, 蔬菜的种类较多, 在某个品种价格较高时, 消费者往往可以选择其它种类蔬菜替代。

(3) 流通因素。随着生活水平的提高, 人们对蔬菜新鲜度的要求越来越高。对于蔬菜来说, 生产地通常是农村或郊区, 消费场所主要在城镇。蔬菜质地鲜嫩, 含水量高, 容易破损腐败, 不适于长期储存, 因此产地和消费地之间的流通十分重要。蔬菜在流通环节的成本对定价也有较大影响。蔬菜物流成本包括高速公路过路费、燃油费、司机劳务费、包装加工费、人工费、场地租赁、产品损耗等。

(4) 综合因素。影响黄瓜价格的综合因素既包括宏观层面上的影响, 也包括微观层面的影响。其中, 宏观层面如国家经济发展水平、人民币汇率、通货膨胀率、国家对蔬菜产业的补助扶持等。微观层面包括黄瓜的季节性、价格周期性

波动趋势、病虫害、极端天气的影响等。

2.2 黄瓜价格波动分析

黄瓜是蔬菜中的主要品类,具有较高的营养价值和食用价值,是居民日常餐桌上的常见食物,产量居中国所有蔬菜品类的前三位^[17]。本研究以2010—2018年北京地区黄瓜价格为例对其价格波动进行分析。黄瓜价格数据主要来源为《中国农产品价格调查年鉴》《中国价格统计年

鉴》、全国农业科教云智能搜索系统 AgSoSo 等。2010—2018 年全国黄瓜月度价格走势如图 1 所示。

从图 1 可以知,黄瓜价格具有很强的季节性,每年呈周期性规律性波动,表现出冬季价高,夏季价低的特点。具体来说,在每年的 12 月至第二年 3 月处于全年峰值,其后持续下跌,往往在 5—7 月达到价格最低点,之后价格呈波动上涨态势。

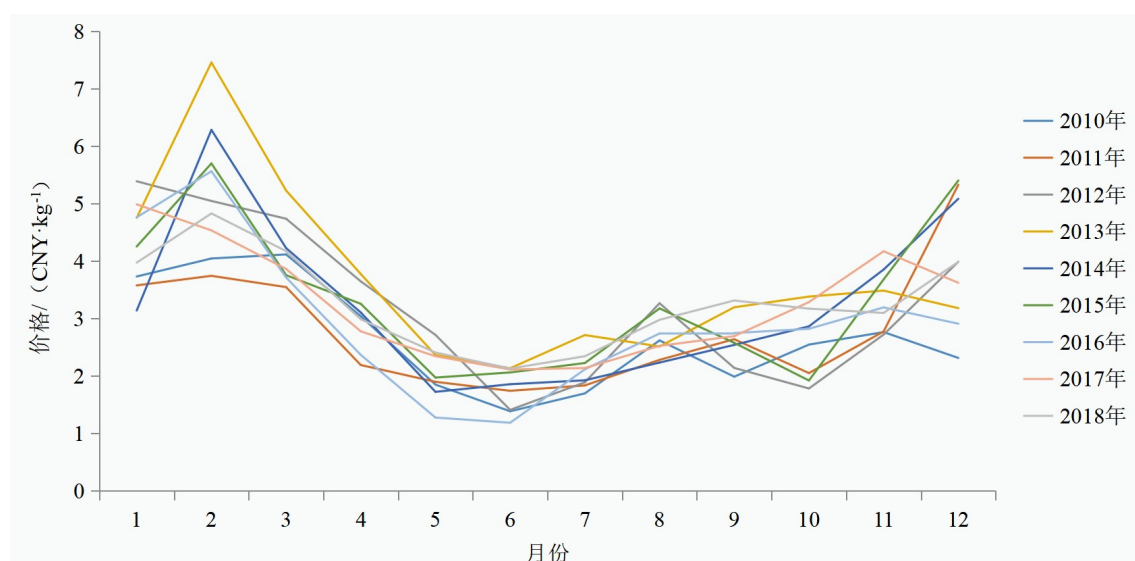


图1 2010—2018年全国黄瓜月度价格走势

Fig. 1 Monthly price trend of cucumbers in China from 2010 to 2018

2.3 黄瓜价格影响因子

根据上述蔬菜价格影响因素分析,结合数据的可获得性,选取以下 24 个因素作为黄瓜价格影响因子。

(1) 供给因素:播种面积、亩产量、受灾面积、土地成本、物质与服务费用和用工成本。

(2) 需求因素:城镇居民人均可支配收入、城镇居民人口数量、城镇蔬菜消费价格指数、叶类蔬菜均价和根茎类蔬菜均价。

(3) 流通因素:高速公路过路费、燃油附加费、司机劳务费、包装加工费、摊位费、人工费和损耗费。

(4) 综合因素:国家经济发展水平、人民币

汇率、通货膨胀率、黄瓜产业扶持金额、气温平均值和气温偏离正常值。

黄瓜价格影响因子数据来源如下:黄瓜播种面积、亩产量、受灾面积来自《中国农业统计资料》;土地成本、物质与服务费用、用工成本、包装加工费、人工费来自《全国农产品成本收益汇编》;黄瓜月价格、叶类蔬菜月均价、根茎类蔬菜月均价来自《中国农产品价格调查年鉴》;城镇居民人均可支配收入、城镇居民人口数量、城镇蔬菜消费价格指数、高速公路过路费、燃油附加费、货运司机劳务费来自于《中国统计年鉴》和《北京统计年鉴》;国家经济发展水平、通货膨胀率、气温平均值变化率、气温偏离正常值来自于 Wind 资讯网站;人民币汇率来自国际

清算银行网站；摊位费、损耗费采用调研数据和文献资料。

3 Lasso 回归与 BP 神经网络结合的组合预测模型构建

3.1 理论基础

(1) Lasso 回归。Lasso (Least absolute shrinkage and selection operator, Tibshirani (1996)) 回归方法^[18]的本质是一种压缩估计。它通过构造一个惩罚函数以得到更加精炼的模型，使得它可以压缩一些系数，同时把一些系数置零，达到子集收缩的效果。

在一般的回归模型中，各个数据的观测值一般认为是彼此独立的，且因为模型中存在许多变量，量纲经常不一样，为排除量纲的干扰，需要对全部自变量 $X_i = (x_{i1}, x_{i2}, \dots, x_{im})$ 做标准化变换，使 $z_{i1}, z_{i2}, \dots, z_{im}$ 的均值为 0，方差为 1，无量纲差别。该回归模型的 Lasso 估计为：

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha, \beta)}{\operatorname{argmin}} \sum_{i=1}^n (y_i - \alpha - \sum_{j=1}^p x_{ij} \beta_j)^2$$

$$\text{subject to } \sum_{j=1}^p |\beta_j| \leq s \quad (1)$$

其中， $s \geq 0$ 是调和参数，对于任意的 s 有 α 的估计 $\hat{\alpha} = \bar{y}$ ，要使减小总体回归系数，只需要对调和参数 s 进行调整。当令 $s^0 = \sum_{j=1}^p |\beta_j|$ ， $s \leq s^0$ 时，某些变量的系数会减小，甚至近似 0 或等于 0，这些无关的或关系极小的自变量将会被筛选，从而使回归模型的精度及可解释性提高。求解 Lasso 回归系数一般使用 Mallows C_p 统计量，这一方法可以评估回归模型的优劣，方法定义如下。

$$C_p = \frac{SSE_p}{\delta^2} - n + 2p \quad (2)$$

$$SSE_p = \sum_{i=1}^n (Y_i - Y_{pi})^2 \quad (3)$$

其中， n 表示样本的总量； p 表示子集回归模型中存在的自变量的数量； SSE_p 表示 p 对因变量 Y 进行回归后的残差的平方和； δ^2 表示所有自变量对因变量 Y 进行回归时，方差的均值的预

测。据此，得到当 C_p 为最小值时的模型，就获得在全局范围内最佳的变量的子集，同时生成的是效果最佳的回归方程。

(2) BP 神经网络。属于前向神经网络，其特点在于加入了反向传播算法，该算法核心是当网络输出的值与实际值有偏差时，将这一差值从后往前传递，每一层网络的权值朝着减小偏差的方向调整，达到全局调整的目的。经过不断地正向反向调整，使偏差降低到预计范围。

3.2 方法结合

以影响因素为基础的预测是根据不同的影响因素与预测变量之间存在的联系，建立神经网络模型为影响因素与预测变量之间建立对应关系，形成预测模型。Lasso 回归方法与 BP 神经网络结合的预测模型 L-BPNN 结构如图 2 所示， $X_1 \dots X_n$ 为初始影响因素， $X_{1*} \dots X_{n*}$ 为 Lasso 回归处理后的影响因素， $h_1 \dots h_m$ 为隐含层的神经元， $w_1 \dots w_m$ 为权重， y_m 为输出的黄瓜价格。

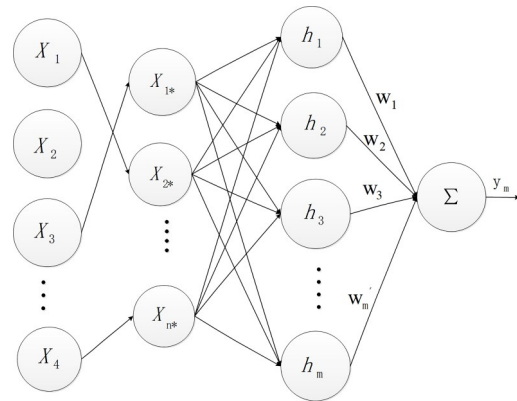


图 2 L-BPNN 结构图

Fig. 2 L-BPNN structure diagram

3.3 数据标准化

L-BPNN 模型输入为上述选取的 24 个主要影响因素，模型的输出为黄瓜的月度价格。为能够更好地体现预测的目的，本研究将选取某自然月的影响因素作为输入，输出则为下一个自然月的黄瓜价格。输入数据做归一化处理，将数据范围控制在 $[0, 1]$ 或 $[-1, 1]$ 之间，预测结束后再

对数据进行反归一化处理,在BP神经网络中使用函数premnmx和postmnmx可以直接将输入数据归一化以及输出数据反归一化。

3.4 基于Lasso回归的黄瓜价格影响因素选取

由于黄瓜的影响因素较多,直接使用BP模型预测存在模型复杂度太高,训练速度慢的问题。本研究将Lasso回归方法与BP神经网络结合,先使用Lasso回归将输入的影响因素压缩,压缩后剩余的关联度较大的影响因素作为BP神经网络的输入的影响因素,再使用神经网络对输入的影响因素与输出的价格进行训练调整模型。

模型的输入为2.3中得出的24个黄瓜价格影响因素,将其全部输入组合模型。根据Lasso回归模型的基本理论以及对黄瓜价格影响因素的分析,建立黄瓜价格影响因素线性回归模型,如公式(3)所示。

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_{23} z_{23} + \beta_{24} z_{24} + \varepsilon \quad (4)$$

其中, Y 表示黄瓜价格; β_0 代表常数项; $\beta_1 \cdots \beta_{24}$ 为各影响因素的系数; ε 为随机误差, $\varepsilon \sim N(0, \delta^2)$ 。

将构建的Lasso回归模型用R语言编程进行实证分析。参数 s 为0时各系数均没有进入模型,随着参数 s 的增大,不断有变量被选入模型,直到所有变量都被选入模型, $s=1$ 时结束,总共使用了35步找出全部解。不同参数 s 的估计方法寻找最优解的结果不同,本研究使用精度与拟合度综合效果较好的赤池信息准则(Akaike Information Criterion, AIC)准则寻找最优解,在AIC达到最小时模型解为最优。表1给出了AIC准则下模型前8步的结果,在第8步时AIC达到了最小,第九步及以后开始上升,此时模型选取的变量共有12个,分别为:播种面积、产业扶持金额、气温平均值、受灾面积、土地成本、城镇居民人均可支配收入、城镇蔬菜消费价格指数、燃油附加费、摊位费、包装加工费、通货膨胀率和气温

偏离正常值。

3.5 L-BPNN神经网络建模

将Lasso回归方法与BP神经网络结合,进行L-BPNN神经网络建模,步骤如下。

(1) 确定网络结构。BP神经网络的结构主要需要确定三个参数,神经网络的层数、隐含神经元的数目和激活函数。

神经网络层数。BP神经网络的输入和输出层均为一层,需要确定的是隐含层的数目。大量理论研究表明^[19,20],多隐含层会极大增加算法复杂度,但收益提升并不是十分明显。因此本研究BP神经网络采用输入层、隐含层、输出层各一层,这也是经典BP神经网络设计结构。

隐含层神经元数量。这个参数是会影响网络性能的重要参数,数量过少会降低预测性能,造成模型不稳定;数量过多则会增加训练时间影响实用性。本研究使用公式(5)来确定神经元数量。

$$n = \sqrt{N + M} + a \quad (5)$$

其中, n 为隐含层神经元数量; N 为输入层样本数量, M 为输出层样本数量; $a \in [1, 10]$; 由输入变量和输出变量可得,本研究神经网络的隐含层神经元数量取值范围为 $[4, 13]$ 。

激活函数。该函数主要用于激活神经元,使之形成合理的输出。激活函数有许多种,本研究采用Sigmoid激活函数。Sigmoid函数被广泛应用于预测,该函数具有良好的非线性映射能力,对于取值有较高的置信度。其计算公式为:

$$f(x_i) = \frac{1}{1 + \exp(-x_i)} \quad (6)$$

其中, x_i 为输入变量。

本研究构建的L-BPNN组合模型输入变量共有24个,其中BP神经网络输入变量12个,由公式(5)可得隐含层节点范围是 $[5, 14]$ 。确定L-BPNN组合模型为四层网络,调用newff函数建立BP神经网络:

$$net = newff(\text{minmax}(\text{inputn}), [5, 1], \{\text{'tansig'}, \text{'purelin'}\}, \text{'trainlm'}) \quad (7)$$

表 1 Lasso 回归模型变量选择与参数估计

Table 1 Variable selection and parameter estimation of Lasso regression model

系数	影响因素	1	2	3	4	5	6	7	8
β_1	播种面积	0	-0.28212	-0.20547	-0.18647	-0.19540	-0.20108	-0.20273	-0.20646
β_2	亩产量	0	0	0	-0.18932	-0.29021	-0.10198	0	0
β_3	受灾面积	0	0	0.408734	0.438949	0.445482	0.564859	0.592184	0.531208
β_4	土地成本	0	0.239479	0.290850	0.303829	0.313982	0.360215	0.347498	0.324921
β_5	物质与服务费用	0	0	0	0.101541	0.107532	0.140198	0.092192	0
β_6	用工成本	0	0	0	0.386323	0.389201	0	0	0
β_7	城镇居民人均可支配收入	0	0	0.523294	0.518713	0.514719	0.590234	0.364021	0.393297
β_8	城镇居民人口数量	0	0.310583	0.210840	0.200987	0.112749	0	0	0
β_9	城镇蔬菜消费价格指数	0	0.412385	0.354817	0.312390	0.378974	0.376213	0.374085	0.375532
β_{10}	叶类蔬菜均价	0	0	0.570197	0.542760	0.423570	0.437502	0	0
β_{11}	根茎类蔬菜均价	0	0.453750	0.421097	0.392375	0.321937	0.291374	0	0
β_{12}	高速公路过路费	0	0	0	0.210870	0.124791	0.021397	0.078730	0
β_{13}	燃油附加费	0	0	0	0.343019	0.341309	0.351035	0.253108	0.271075
β_{14}	司机劳务费	0	0	0	0.039218	0.032957	0.023985	0.093213	0
β_{15}	摊位费	0	0	0	0.196657	0.207866	0.251224	0.243317	0.213077
β_{16}	人工费	0	0	0	0.217553	0.233591	0	0	0
β_{17}	损耗费	0	0.353411	0.328929	0.293665	0.227755	0.236801	0	0
β_{18}	包装加工费	0	0	0.217917	0.397591	0.319327	0.343109	0.327519	0.247327
β_{19}	国家经济发展水平	0	0	0.153017	0.043092	0.039759	0.043197	0.022174	0
β_{20}	人民币汇率	0	0	0.012874	0.023857	0	0	0	0
β_{21}	通货膨胀率	0	0	0.565325	0.610128	0.540198	0.531208	0.531093	0.501098
β_{22}	黄瓜产业扶持金额	0	0	0	0	-0.320180	-0.290180	-0.341820	-0.351080
β_{23}	气温平均值	0	-0.43296	-0.532850	-0.624810	-0.619750	-0.603290	-0.593820	-0.592190
β_{24}	气温偏离正常值	0	0.431028	0.413098	0.323093	0.413902	0.532101	0.632181	0.491273
AIC			3.228796	1.967890	0.891344	0.245776	0.102355	0.083123	0.071226

其中，隐含层节点从 5 开始，随训练过程增加或调整节点数量，输出层节点为 1，隐含层的激活函数为 tansig，输出层的激活函数为 purelin，训练函数为 trainlm。

(2) 网络训练。神经网络的基本训练流程如下。

① 初始化权值。

$$\omega_{sp} = \text{Random}(\cdot) \quad (8)$$

其中， sp 为 ij 、 jk 、 kl 。

输入 p 个样本，且 p 为当前输入样本。

② 得到各层节点的输出 x_j' 、 x_k' 及 y_i 。

③ 计算各层反传的偏差。

$$\delta_{kl}^{(p)} = (d_l^{(p)} - y_l^{(p)}) y_l^{(p)} (1 - y_l^{(p)}) \quad (9)$$

$$\delta_{jk}^{(p)} = \sum_{l=0}^{m-1} \delta_{kl}^{(p)} \omega_{kl}'' x_k^{(p)} (1 - x_k^{(p)}) \quad (10)$$

$$\delta_{ij}^{(p)} = \sum_{k=0}^{n_2} \delta_{jk}^{(p)} \omega_{jk}' x_j^{(p)} (1 - x_j^{(p)}) \quad (11)$$

其中， $j=0, 1, 2, \dots, n_1$ ； $k=0, 1, 2, \dots, n_2$ ； $l=0, 1, 2, \dots, m-1$ 。

④ 循环网络直至 p 个样本都训练完。

⑤ 根据权值调整公式 (11) 对各层节点的权值进行调整直到每个 p 和 l 都满足 $|d_l^{(p)} - y_l^{(p)}| < \varepsilon$ 或训练次数达到上限，训练结束。

$$\begin{aligned}\omega_{ij}(t+1) &= \omega_{ij}(t) + \eta \sum_{p=1}^p \delta_{ij}^{(p)} x_i^{(p)} \delta_{ij}^{(p)} \\ &= \sum_{k=0}^{n_2} \delta_{jk}^{(p)} \omega'_{jk} x_j^{(p)} (1 - x_j^{(p)})\end{aligned}\quad (12)$$

调用train函数对神经网络进行训练,神经网络的训练总次数为10,000次,目标误差为0.00001,只要达到其中之一条件则结束。调用sim函数对训练完成的模型进行测试。

(3) 仿真实验。训练好的神经网络需要进行仿真,通常采用sim函数进行,其格式如下。

$$Y = \text{sim}(\text{net}, P) \quad (13)$$

其中, Y 表示神经网络的输出; net 表示训练完成的模型; P 为网络的输入数据矩阵。

4 结果与分析

4.1 L-BPNN 组合模型预测结果分析

以2010年1月至2018年4月的影响因素与2010年2月至2018年5月的黄瓜价格数据作为训练数据,反复调整隐含层节点数,当节点数小于11时,预测相对误差超过5%,节点数大于11时,每增加一个节点,训练时间增加40%,发现隐含层节点数为11时网络训练结果最佳。模型训练完成后将2018年5月至11月的价格影响因素输入模型,将预测的2018年6月至12月的结果与真实值进行比较,如表2所示。

由表2可得,有5个月份的预测相对误差低于1%,剩余2个月份的预测绝对误差介于1%与

表2 L-BPNN 组合模型预测结果

Table 2 Prediction results of L-BPNN combined model

月份	真实值/CNY	预测值/CNY	绝对误差/CNY	相对误差/%
6	2.129618364	2.127834559	-0.001783805	0.08
7	2.343750000	2.351965774	0.008215774	0.35
8	2.978162865	2.929674385	-0.048488480	1.62
9	3.314072948	3.324556398	0.010483450	0.31
10	3.173002850	3.195643321	0.022640471	0.71
11	3.097538197	3.092345332	-0.005192865	0.17
12	3.987193719	3.931245568	-0.055948151	1.40

2%之间,总体平均相对误差仅为0.66%,低于5%的模型检验标准,说明模型通过了实际价格检验。具体来看,预测值低于真实值的月份有4个月,预测值高于真实值的月份有3个月,说明预测值没有明显的偏向。从价格趋势来看,有4个月份在价格变化趋势上大于真实值的变化率,3个月份在价格趋势上小于真实值的变化率,说明变化趋势没有明显的影响。该预测模型取得了良好的预测效果,唯一不足之处在于模型复杂,训练时间较长,且多次陷入局部极小值,对参数及隐含层神经元进行调整复杂度较大。

4.2 L-BPNN 与其它方法实验结果对比

将基于影响因素的L-BPNN组合模型与基于影响因素的Lasso回归模型、基于时间序列的BP神经网络模型、基于时间序列的RBF神经网络模型这四组对照组模型的预测结果进行对比,结果如图3所示。

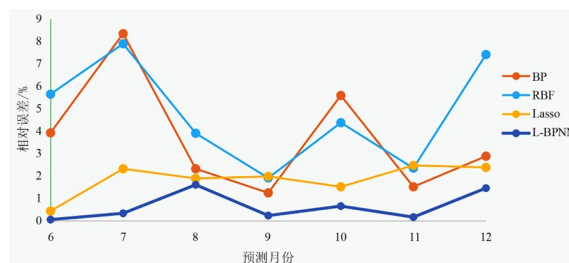


图3 L-BPNN 模型与其它3种模型预测结果

相对误差对比

Fig. 3 Comparison of the relative error of the prediction results between the L-BPNN model and the other three models

总体来看,平均相对误差最小的是L-BPNN组合模型,为0.66%,低于1%;其次为Lasso回归模型,为1.86%,介于1%与2%之间;最后是BP神经网络与RBF神经网络预测效果则较差,平均相对误差分别是3.69%和4.78%,介于3%到5%之间。可见在本研究的黄瓜价格预测中,L-BPNN组合模型在基于黄瓜影响因素的预测中有较好的表现。

从相对误差的最小值来看,L-BPNN组合模型的最小相对误差为0.08%,分别比Lasso回归

模型、BP 神经网络模型、RBF 神经网络模型的最小相对误差低 81.82、93.6、95.81 个百分点。从相对误差的最大值来看, BP 神经网络模型的最大相对误差为 8.33, 比 L-BPNN 组合模型、Lasso 回归模型、RBF 神经网络模型的最大相对误差高 414.2%、237.25% 和 5.71%。平均误差最小的 L-BPNN 组合模型, 分别比 Lasso 回归模型、BP 神经网络模型、RBF 神经网络模型的平均相对误差低 64.52%、82.11% 和 86.2%。可见 L-BPNN 组合模型较其他 3 种模型有较明显的误差减小, 精度提升。

5 结论

本研究首先从供给、需求、流通等方面分析了影响蔬菜价格的因素, 并以黄瓜为研究对象, 确定了 24 个黄瓜价格的影响因子。然后运用 Lasso 回归方法对黄瓜价格影响因素进行定量分析, 利用 Lasso 回归压缩参数的特性, 将影响因素按照关联度排序, 剔除影响较小的因素, 筛选得到 12 个关联度较大的影响因素。在此基础上, 构建了 Lasso 回归方法与 BP 神经网络相结合的组合模型 (L-BPNN), 建立黄瓜价格与影响因素的对应关系, 通过已知的影响因素预测未来价格, 对黄瓜价格开展短期预测。本研究取得的主要结论如下。

(1) 12 项关联度较大的黄瓜价格影响因素包括: 正向影响有土地成本、城镇居民人均可支配收入、城镇蔬菜消费价格指数、燃油附加费、摊位费、包装加工费、通货膨胀率、受灾面积、气温偏离正常值; 负向影响有播种面积、产业扶持金额、气温平均值。

(2) 黄瓜价格预测结果。以 2010 年 2 月至 2018 年 5 月北京地区黄瓜价格数据为例, 将 L-BPNN 预测结果分别与基于时间序列的 BP 神经网络、RBF 神经网络以及基于影响因素的 Lasso 回归模型的预测结果进行比较, L-BPNN 组合模型具有相对较高的预测精度, 其平均相对误差仅为 0.66%, 组合模型与单一的基于影响因素的预

测、基于时间序列的预测方法等相比, 有较好的表现。

未来的研究中, 一方面考虑加入更多的价格影响因素, 如种子肥料农药等生产资料成本、装车费等; 另一方面, 可以适当加大数据样本量和数据精细度, 如增加更多历史月份价格数据或日价格数据, 以期提高模型训练的效果, 提高价格预测精度。

参考文献:

- [1] 曹宏鑫, 葛道阔, 张文字, 等. 农业模型发展分析及应用案例[J]. 智慧农业(中英文), 2020, 2(1): 147-162.
CAO H, GE D, ZHANG W, et al. Developmental analysis and application examples for agricultural models[J]. Smart Agriculture, 2020, 2(1): 147-162.
- [2] 孔繁涛, 朱孟帅, 孙坦. 现代信息技术在农业领域的应用分析与建议——互联网企业进军农业引发的思考[J]. 智慧农业, 2019, 1(4): 31-41.
KONG F, ZHU M, SUN T. Application analysis and suggestions of modern information technology in agriculture: Thoughts on Internet enterprises entering agriculture[J]. Smart Agriculture, 2019, 1(4): 31-41.
- [3] 许世卫, 李哲敏, 李干琼, 等. 农产品市场价格短期预测研究进展[J]. 中国农业科学, 2011, 44(17): 3666-3675.
XU S, LI Z, LI G, et al. Advances in research of short-term forecasting methods of agricultural product price[J]. Scientia Agricultura Sinica, 2011, 44(17): 3666-3675.
- [4] AHINLI M A. Potato price forecasting with Holt-Winters and ARIMA methods: A case study[J]. American Merican Journal of Potato Research, 2020, 97(4): 336-346.
- [5] GHOSH S, SINGH K N, THANGASAMY A, et al. Forecasting of onion (*Allium cepa*) price and volatility movements using ARIMAX-GARCH and DCC models[J]. Indian Journal of Agricultural Sciences, 2020, 90(5): 169-173.
- [6] PAREDES-GARCIA W J, OCAMPO-VELAZQUEZ R V, TORRES-PACHECO I, et al. Price forecasting and span commercialization opportunities for mexican agricultural products[J]. Agronomy-basel, 2019, 9(12): 826-828.

- [7] 陈芳, 楼文高. 基于广义回归神经网络的蔬菜市场日价格预测[J]. 浙江农业学报, 2015, 27(7): 1253-1258.
CHEN F, LOU W. Forecasting of vegetable daily price based on general regression neural network[J]. Acta Agriculture Zhejiangensis, 2015, 27(7): 1253-1258.
- [8] 叶露, 李玉萍, 秦小立, 等. 基于 PSO-BP 与 RBF 神经网络的蔬菜价格组合预测[J]. 北方园艺, 2015(21): 212-215.
YE L, LI Y, QIN X, et al. Vegetables price combination forecasting based on PSO-BP and RBF neural network[J]. Northern Horticulture, 2015(21): 212-215.
- [9] 刘合兵, 韩晶晶, 席磊. 小波变换-BP 神经网络的农产品价格预测研究[J]. 中国农业信息, 2019, 31(6): 85-92.
LIU H, HAN J, XI L. Agricultural product price forecast based on wavelet transform and BP neural network[J]. China Agricultural Informatics, 2019, 31(6): 85-92.
- [10] 贾宁, 郑纯军. 基于 LSTM-DA 神经网络的农产品价格指数短期预测模型[J]. 计算机科学, 2019, 46(S2): 62-65, 71.
JIA N, ZHENG C. Short-term forecasting model of agricultural product price index based on LSTM-DA neural network[J]. Computer Science, 2019, 46(S2): 62-65, 71.
- [11] XIONG T, LI C, NBAO Y. Seasonal forecasting of agricultural commodity price using a hybrid STL and ELM method: Evidence from the vegetable market in China[J]. Neurocomputing, 2018(275): 2831-2844.
- [12] ZHANG D, ZANG G, LI J, et al. Prediction of soybean price in China using QR-RBF neural network model[J]. Computers and Electronics in Agriculture, 2018, 154:10-17.
- [13] 张桂喜, 马立平. 预测与决策概述[M]. 北京: 首都经济贸易大学出版社, 2006.
ZHANG G, MA L. An introduction to forecast and decision[M]. Beijing: Capital University of Economics Press, 2006.
- [14] 许世卫, 张峭, 李志强, 等. 番茄价格形成及利润分配调查报告[J]. 农业展望, 2008(5): 3-5.
XU S, ZHANG Q, LI Z, et al. Investigation report on tomato price formation and profit distribution [J]. Agricultural Outlook, 2008(5): 3-5.
- [15] 徐依婷, 穆月英, 赵友森. 北京市蔬菜价格变动及影响因素分析[J]. 中国蔬菜, 2017(11): 62-69.
XU Y, MU Y, ZHAO Y. Vegetable price fluctuation in Beijing and analysis of its influence factors[J]. China Vegetables, 2017(11): 62-69.
- [16] 黄靖贵, 李春红. 基于空间面板数据模型的蔬菜价格影响因素研究[J]. 广西大学学报(自然科学版), 2019, 44(6): 1842-1850.
HUANG J, LI C. Study on the effects of vegetable price based on spatial panel data model[J]. Journal of Guangxi University(Nat Sci Ed), 2019, 44(6): 1842-1850.
- [17] 张晶, 吴建寨, 孔繁涛, 等. 2019 年我国蔬菜市场运行分析与 2020 年展望[J]. 中国蔬菜, 2020(1): 1-8.
ZHANG J, WU J, KONG F, et al. Analysis of China's vegetable market operation in 2019 and prospect in 2020[J]. China Vegetables, 2020(1): 1-8.
- [18] TIBSHIRANI R. Regression shrinkage and selection via the lasso: A retrospective[J]. Journal of the Royal Statistical Society: Series B Statistical Methodology, 2011, 73(3): 273-282.
- [19] 潘文婵, 刘尚东. BP 神经网络的优化研究与应用[J]. 计算机技术与发展, 2019, 29(5): 74-76, 101.
PAN W, LIU S. Optimization research and application of BP neural network[J]. Computer Technology and Development, 2019, 29(5): 74-76, 101.
- [20] 董志贵, 王福林, 宋庆凤, 等. 基于 BP 神经网络的无约束优化方法[J]. 统计与决策, 2019, 35(1): 79-82.
DONG Z, WANG F, SONG Q, et al. Unconstrained optimization method based on BP neural network[J]. Statistics and Decision, 2019, 35(1): 79-82.

Short-Term Price Forecast of Vegetables Based on Combination Model of Lasso Regression Method and BP Neural Network

YU Weige^{1,2}, WU Huarui^{1,2,3*}, PENG Cheng^{1,2,3}

(1. *Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China;*

2. *National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China;*

3. *Key Laboratory of Agri-informatics, Ministry of Agriculture and Rural Affairs, Beijing 100097, China*)

Abstract: Vegetables are an important part of residents' diet. The abnormal fluctuation of vegetable prices has caused losses to the economic interests of vegetable farmers and also affected the daily diet and quality of life of residents. However, there are some difficulties in vegetable price prediction, such as large price fluctuation and complicated influencing factors. Cucumber is the main category of vegetables and a common food on the daily table of residents and its recent price fluctuations have aroused widespread concern. In this research, taking cucumber as the research object, a combination model (L-BPNN) combining Lasso regression method and BP neural network was constructed to forecast the short-term price of cucumber. Firstly, the factors affecting the price of cucumber, such as supply, demand and circulation were analyzed. Then the price fluctuation characteristics of cucumber in China from 2010 to 2018 were analyzed and 24 factors were selected as the influencing factors of cucumber price. In the case of complex factors, Lasso regression was used to compress the 24 input influencing factors and the 12 remaining influencing factors with large correlation degree after compression were used as the input influencing factors of BP neural network. Among the 12 related factors, the positive effects included: land cost, per capita disposable income of urban residents, urban vegetable consumption price index, fuel surcharge, booth fee, packaging and processing fee, inflation rate, affected area and temperature deviation from normal value; negative effects included sown area, industrial support amount and average temperature. On this basis, a combination model combining Lasso regression method with BP neural network (L-BPNN) was constructed to forecast the short-term price of cucumber. The neural network was used to train and adjust the model between the input influencing factors and the output price. Compared with the regression analysis and intelligent analysis methods, the results show that the average relative error of L-BPNN combination model was the smallest, only 0.66%, which was 64.52%, 82.11% and 86.2% lower than Lasso regression model, BP neural network model and RBF neural network model respectively, and had higher prediction accuracy. The results of this study realizes the short-term price forecast of cucumber, and can also be extended to other vegetable varieties, which is of great significance for guaranteeing the income of vegetable farmers and stabilizing the market price of vegetables.

Key words: vegetable; influencing factors; price forecast; combination model; Lasso regression; BP neural network; RBF neural network

(登陆 www.smartag.net.cn 免费获取电子版全文)