

The Price Prediction of Vegetables by Using Machine Learning

Qiaochu Chen ^{1,*}, Zhihua Yu ² and Kaixin Zheng ³

¹ School of Information and Communication Engineering, Beijing University of Post and Telecommunication, Beijing, China

² One Direction Academy, Ontario, Canada

³ Sino-Canada School Suzhou, Suzhou, China

* Corresponding author: chenqiaochu@bupt.edu.cn

Abstract. Vegetable prices are great important in ensuring people's livelihood. With the progress of science and technology, machine learning algorithms have been applied in the field of vegetable price prediction. This paper analyzed the model that has been used in vegetable price prediction and its results and compared the accuracy of each model. The study in this paper concluded that LSTM is suitable for handling sequence data with long-term dependence but this model requires a large amount of data for training. ARIMA is suitable for short-term predictions but may not fit well with non-stationary and nonlinear data. In addition, this paper introduces other algorithms. For example, the paper shows How to use Lasso regression to compress the input variables and how to reduce model complexity and improve model generalization ability. The BP neural network was used to fit the output variables to improve the model fitting ability. This paper further studies and summarizes the combination of each model and the accuracy of the combined model for predicting vegetable prices. It can be seen that the prediction error of the L-LBNN model is very small, which can be effectively used in the price prediction of vegetables. The purpose of this study is to summarize the current situation of predicting vegetable prices with machine learning algorithms to provide guidance and inspiration for researchers and to promote the development of the use of machine learning in the field of price prediction of vegetables.

Keywords: Machine Learning, Prediction, Vegetable.

1. Introduction

With the development of the social economy, vegetables, as an important part of residents' diets, have a significant impact on the income of vegetable farmers, the operation of supermarkets, the welfare of consumers, etc. Therefore, predicting vegetable prices is of great importance for ensuring market supply and demand balance, stabilizing market prices, improving economic efficiency, etc. However, vegetable prices are affected by various factors, such as supply, demand, circulation, season, weather, policy, etc., and their price fluctuations are large and difficult to predict. Traditional prediction methods, such as regression analysis, time series analysis, etc., often cannot fully utilize the information of data and have low prediction accuracy. Therefore, how to use advanced machine learning techniques to improve the accuracy and effectiveness of vegetable price prediction is a worthwhile research problem. In recent years, machine learning, as an important branch of artificial intelligence, has shown strong application capabilities in various fields. Machine learning refers to the process of letting computers learn rules and knowledge from data, to achieve intelligence. Machine learning methods can be divided into supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, etc. Among them, supervised learning refers to using known input and output data to train a model that can predict unknown output [1].

Based on the existing research, this paper mainly explores the different prediction results and accuracy of various algorithms for vegetable price prediction. This paper aims to provide a comprehensive analysis of machine learning for short-term prediction of vegetable prices.

This paper reviews the relevant research on vegetable price prediction at home and abroad, and points out the applicability of machine learning methods in vegetable price prediction. In addition, this paper mainly explores the specific application of LSTM, L-BPNN, Lasso regression and BP

neural network in vegetable price prediction. Through this study, we can use the actual vegetable price data to build and validate different machine learning models, evaluate their prediction accuracy and stability, and provide reliable tools and methods for vegetable price prediction.

2. Necessity Analysis of Price Prediction

Price prediction can help the government formulate reasonable macroeconomic policies, such as monetary policy, fiscal policy, industrial policy, etc., to promote economic growth, stabilize prices, balance trade, adjust income distribution, and other goals. For example, through price prediction, the government can adjust policy tools such as interest rates, exchange rates, tax rates, subsidies, etc. in advance, to influence market supply and demand and expectations, and thus achieve the management of aggregate demand and aggregate supply.

2.1. Significance of vegetable price forecasting

Food is the paramount necessity of the people, and the prediction of vegetable prices will have a considerable impact on the vegetable market and even the entire video market, which is reflected in the following aspects. In terms of guiding supply adjustment, the current vegetable market is mainly regulated by the market, and predicting vegetable prices in advance can guide vegetable producers and importers to adjust the number of vegetables flowing to the market, which can help suppliers avoid risks and promote the healthy development of the market while stabilizing prices and reducing waste. In addition, price prediction will affect consumers' purchase decisions. Predicting vegetable prices helps to eliminate the information gap between consumers and suppliers, helps to disperse demand, and stabilizes the market. In terms of guiding agricultural investment. Long-term vegetable price prediction helps investors discover potential investment opportunities in the vegetable market, which is conducive to attracting capital and expanding the scale of the market. Finally, the corresponding prediction can provide a basis for policy intervention. The current vegetable prices are mainly regulated by the market, and the spontaneity of the market economy can maintain the balance of market supply and demand in a longer time scale, but its blindness and lag will inevitably increase the price fluctuations when facing market environment changes, and it cannot guarantee that the formed supply and demand relationship is reasonable and healthy. Therefore, policy intervention in the market is necessary. Predicting prices helps the government and market regulators obtain more accurate information in advance, reserve time for policy formulation and implementation, and enable them to intervene in a timely manner to ensure the healthy development of the market.

2.2. The role of machine learning in vegetable price prediction

The price of vegetables is influenced by a variety of factors. Natural factors include the climate, light and terrain of the production area, while human factors include transportation and storage conditions, market supply and demand, and policy formulation. In addition, there are a large number of hidden factors, each of which has a very complex and difficult-to-quantify cause. Vegetable markets are common in various places, with a wide variety of types, and different markets are affected by different factors. This complexity makes it difficult to apply traditional algorithms, and manual analysis makes it difficult to achieve comprehensiveness and accuracy. In this context, data related to vegetable prices are relatively easy to collect. Machine learning, as an automated and intelligent algorithm, has powerful performance in dealing with complex problems. With sufficient data support, applying machine learning to vegetable price prediction can save a lot of labor costs and achieve higher accuracy. Therefore, the application of machine learning in vegetable price prediction can be understood as using historical data and influencing factors to establish a mathematical model that accurately reflects the law of vegetable price changes, and using this model to predict and analyze future vegetable prices. Machine learning algorithms also have a certain generalization and scalability, and can cope with various diversified prediction needs. Therefore, machine learning plays a very significant role in the field of vegetable price prediction. According to different machine learning

algorithms, different types of vegetable price prediction models can be constructed, such as linear regression models, decision tree models, neural network models, etc. These models have their own advantages and disadvantages and need to choose the appropriate model according to the actual situation. For example, linear regression model is a simple and commonly used model, which assumes that there is a linear relationship between vegetable prices and influencing factors, and solves the model parameters by minimizing the sum of squared errors. However, the linear regression model may not be able to handle nonlinear and high-dimensional data well, so it needs to preprocess or select features for the data. The decision tree model is a model based on tree structure, which divides the data into different subsets and makes predictions on each subset. Decision tree model can handle nonlinear and categorical data, but it may have overfitting or underfitting problems, so it needs to control the depth and complexity of the tree. Neural network model is a model based on artificial neurons, which can fit complex function relationships through multi-layer nonlinear transformations. Neural network model can handle high-dimensional and nonlinear data, but it requires a lot of data and computing resources, and may have local optimum or gradient disappearance problems, so it needs to optimize the network structure and parameters. Machine learning has an important significance for vegetable price prediction, which can help vegetable producers, operators and consumers understand market dynamics, formulate reasonable production, sales and purchase strategies, and ensure the stability and development of vegetable market.

3. Analysis of Machine Learning Algorithm in Vegetable Price Prediction

3.1. LSTM and ARIMA

3.1.1. LSTM

LSTM is a deep learning model, which is a variant of recurrent neural network (RNN), with memory cells and gating mechanisms, which can capture long-term dependencies, and usually has better prediction ability when dealing with sequence data with longer time dependencies. LSTM requires relatively less data, and can handle non-stationary sequences, but as a deep learning model, it usually takes a long time to train when the data volume is large [1].

LSTM is a specific form of RNN, which has a significant drawback called long term dependency. When the input is a time series, the RNN model processes the input of one unit and the output of the previous unit at a time, and outputs a single prediction. Therefore, the output of distant units will gradually disappear due to multiple processing, which is the long-term dependency. LSTM model improves and solves the problem of RNN model, and adds weights and biases to retain key information as much as possible and remove unnecessary information. Fig. 1 shows the operation process of LSTM model.

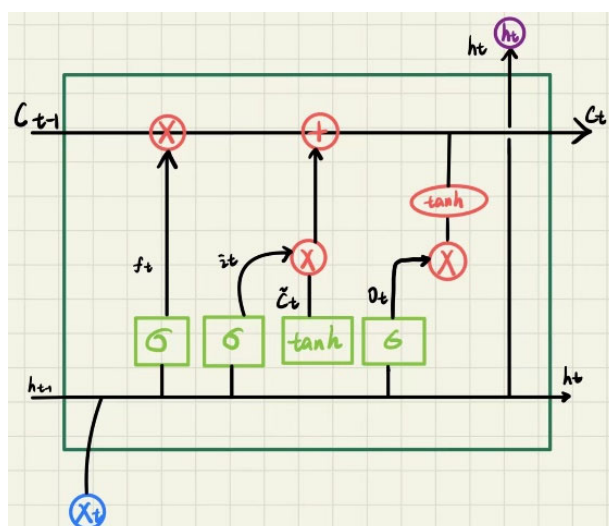


Figure 1. The operation process of LSTM model

3.1.2. ARIMA

The basic idea of ARIMA model is to use the historical information of the data itself to predict the future. The label value at a time point is influenced by both the label values in the past period and the accidental events in the past period, which means that the ARIMA model assumes that the label value fluctuates around the big trend of time, where the trend is influenced by the historical labels, and the fluctuation is influenced by the accidental events in a period of time, and the big trend itself may not be stable. ARIMA requires strict data requirements, requiring that the time series data is stationary (mean and variance do not change with time), so it needs to be preprocessed by splitting and other operations. ARIMA is suitable for short-term prediction tasks, but usually only performs well on linear problems. As a traditional statistical model, the training time is relatively short. ARIMA model consists of three parts, namely autoregressive model (AR), difference process (I) and moving average model (MA). In ARIMA model, the AR part is used to deal with the autoregressive part of the time series, considering the influence of several periods of observation values on the current value (for example, the influence of yesterday, the day before yesterday, etc. on the price of vegetables today). The I part is used to make the non-stationary time series reach a steady state, and can eliminate the trend and seasonal factors in the time series by first-order or second-order splitting. The MA part is used to deal with the moving part of the time series, considering the influence of past prediction errors on the current value (for example, the abnormal rise of local apple prices due to the burning of a certain apple planting area one day). By combining these three parts, ARIMA model can capture the trend changes of the data, and also deal with those data that have temporary and sudden changes.

3.1.3. Comparative analysis

Some problems of LSTM in the application of vegetable price prediction project: when predicting a single type of vegetable, the data volume is often small, and it is difficult to obtain the complete time series of vegetable price changes. For such situations with small data volume, LSTM model is prone to overfitting. LSTM requires a lot of data, because LSTM is a model obtained by repeated repetition, which is a long-term memory cycle for sequential and nonlinear data. The disadvantage of ARIMA model in vegetable price prediction is that it has weak long-term prediction ability. Because ARIMA model only considers the lag and difference of the data, and does not consider other factors that may affect the data, which makes it perform well in short-term prediction of vegetable prices. But when you want to use this model to predict the long-term range of vegetable price fluctuations, there will be a large error. In addition, ARIMA model may not fit well with non-stationary and nonlinear data, and need to process and transform the data appropriately.

In general, when the original data fluctuates little, ARIMA is good, because the prediction results will be linked to the historical results (the historical results fluctuate little, the current fluctuation is also small, naturally the prediction is accurate). It performs well on linear problems, but often has large errors on nonlinear problems. When the fluctuation is large, LSTM is good. But for the situation of small fluctuation and small data volume, LSTM is easy to be too advanced, and overfitting will occur (the error on the training data set decreases in the subsequent epochs, but the error on the validation set is not the case, it reaches the minimum in the second epoch, and then fluctuates). It has high requirements for the data set (because it is a model obtained by repeated repetition, it performs long-term memory cycle on sequential and nonlinear data).

3.2. Lasso, BP and RBF

3.2.1. Principles and Concepts

Least Absolute Selection and Shrinkage Operator is a shrinkage estimation method, which constructs a penalty function, which can compress the coefficients of variables and make some regression coefficients zero, and thus achieve the purpose of variable selection. This method performs variable screening and complexity adjustment while fitting generalized linear models. The main advantage of Lasso regression is that it is especially suitable for the situation where the feature dimension n is much higher than the sample number m . The worst computational complexity of the

algorithm is similar to the least squares method, but its computation speed is almost the same as the forward selection algorithm. It can produce the complete path of piecewise linear results, which is very useful in model cross-validation. The main disadvantage is that since the iteration direction of Lasso is determined by the residual of the target, the algorithm is very sensitive to the noise of the sample. In practical applications, Lasso regression is often used to screen and select related genes after preselection and construct prognostic models.

BP neural network belongs to forward neural network, which is characterized by adding back propagation algorithm. The core of this algorithm is that when the value output by the network is biased from the actual value, this difference is transmitted from back to front, and the weight of each layer of network is adjusted in the direction of reducing the bias, so as to achieve the purpose of global adjustment [2]. The advantage of BP neural network lies in its nonlinear mapping ability and flexible network structure. The number of hidden layers and units in each layer can be changed according to the specific situation. This makes it have excellent adaptability to complex situations. But BP neural network also has some defects. For example, the learning speed of the algorithm is slow, and the complexity of its structure may cause BP neural network to still need a lot of training time when facing simple problems. If the learning rate is set too small, it may fall into the local minimum and fail to achieve the best accuracy. The algorithm also has the limitations of difficult to grasp the optimal number of network layers and neurons, and insufficient generalization ability.

RBF neural network uses radial basis function as the activation function. Compared with BP neural network using gradient descent method, it solves the parameters of mirror base function to approximate the minimum error by determining the center vector and width parameter of hidden layer unit. RBF function also has the characteristic of global approximation, which makes it not easy to fall into the minimum value, and its efficiency is higher and the learning speed is faster. However, RBF neural network also has some disadvantages. The training of RBF neural network involves solving the parameters of radial basis function, which may take more training time. At the same time, when RBF neural network and BP neural network deal with the same problem with the same accuracy requirement, the structure of RBF neural network is more complex, which will cost more hardware resources.

3.2.2. Combination of Lasso and BP neural network

In practical operation, the training conditions are often not ideal, such as insufficient data volume, long training time, and reduced timeliness of prediction results. In this context, measures need to be taken to improve the training efficiency in practical operation. One idea is to preprocess the input data. According to the previous text, there are many factors affecting vegetable prices, and directly using neural networks will make the network structure too complex. Preprocessing methods can be used to compress the influencing factors and extract data with higher correlation as input, so as to simplify the network structure. For example, Lasso regression and BP neural network are combined to form L-BPNN neural network. In the specific implementation process, some factors that may be related to vegetable prices need to be manually screened, and their data need to be statistically analyzed [3, 4]. Use Lasso regression model and normalized statistical data to establish a linear regression model of vegetable price influencing factors, such as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (1)$$

Where y represents the price of vegetables, β_0 Representative constant term $\beta_1 \sim \beta_n$ Is the coefficient of each influencing factor, ε is the random error, $\varepsilon \sim N(0, \delta^2)$. In this process, the strength of the penalty term needs to be constantly adjusted to ensure the optimal fitting accuracy. In this process, the strength of the penalty term needs to be constantly adjusted to ensure the optimal fitting accuracy. The fitting accuracy is usually judged by the Akaike Information Criterion, and when the AIC is the smallest, the factors with higher correlation can be screened out. By bringing the screened factors into the BP neural network, the complexity of the model can be greatly reduced while maintaining the accuracy

3.2.3. PSO-BP neural network

PSO-BP is an optimization method that combines particle swarm optimization (PSO) and BP neural network. The implementation method of PSO is as follows. First, set the running parameters of PSO algorithm, randomly initialize the particle position and particle velocity, determine the particle swarm size, the dimension of individual particles, and map the particles to the corresponding weights and thresholds. Then, according to the mean error fitness function between the measured value and the predicted value, calculate the fitness value of each initial particle, which is the mean square error between the predicted output and the expected output, and find the individual extreme and group extreme. Finally, update the velocity and position of each individual particle, recalculate the particle fitness value, and update the individual extreme and group extreme according to the fitness value of the new particle, until the condition is met, otherwise continue to update the individual. In the process of combining with BP neural network, it is necessary to set up BP neural network first, run PSO algorithm, and use the optimal individual position as the weight and threshold of BP neural network, and then train it [5]. This operation can improve the generalization ability of the model and help to obtain the global optimal solution.

3.3. Hybrid Model of ARIMA and LSTM

There are many irregular factors in the market and the seasonal interference to the model prediction is significant. Hence the study considers eliminating the seasonal influence and combining two models.

3.3.1. Take seasonal factors into consideration:

Vegetable prices are subject to seasonal factors. Simple processing of raw data without considering seasonal factors will lead to deviation [6].

Season index:

First-quarter seasonal Index:

$$SI_1 = \sum_1^n SI_{4(j-1)+1} \quad (2)$$

Second-quarter seasonal Index:

$$SI_2 = \sum_1^n SI_{4(j-1)+2} \quad (3)$$

Third-quarter seasonal Index:

$$SI_3 = \sum_1^n SI_{4(j-1)+3} \quad (4)$$

Fourth-quarter seasonal Index:

$$SI_4 = \sum_1^n SI_{4(j-1)+4} \quad (5)$$

3.3.2. The combination of ARIMA and LSTM models

Using seasonal index: Decomposed the raw vegetable price sequence $Y(t)$ by using the seasonal index-adjusted model to obtain a new data sequence. Then model the data sequence by using ARIMA.

During modeling, it is necessary to analyze the value of APF of all the sequences, to determine the values of p , q , d (ARIMA parameter), to estimate the correlation coefficient of AR and MA, and to test for the correlation of the white noise sequence so that the prediction model can be completed [7].

Because the experimental data of vegetable price is relatively small in the LSTM model, it is necessary to debug the experiment to determine the most appropriate experimental structure.

Generally, we use the Relu function for nonlinear treatment, use Adam function for optimization and MSE use as the loss function.

First predicting the original price sequence with ARIMA to obtain the prediction value f_t and subtracting the true value from f_t to obtain the error sequence. Then predicting the error sequence with the LSTM model to obtain the prediction value s_t . The predictive value of the combined model is $Y_t = f_t + s_t$ [6].

The results found that the average prediction error of is 15.7% for LSTM model, 21.1% for ARIMA model and 11.3% for the combination of ARIMA and LSTM, which is better than the two separate models.

4. Conclusion

This paper aims to summarize existing research results, conduct an overall analysis of the field of vegetable price prediction and the machine learning algorithms used in it, and draw conclusions about the value of using machine learning algorithms to predict vegetable prices and the advantages and disadvantages of each algorithm. This article first studies the value of predicting vegetable prices and finds that predicting vegetable prices can reduce uncertainty, facilitate producers, operators, consumers, and market supervision departments to make more rational decisions, and greatly promote social stability and harmony. The multivariate nature of vegetable price influencing factors is very favorable for the application of machine learning. In the survey process, this article finds that LSTM, ARIMA, Lasso, BP, and RBF are often used methods. This article analyzes their principles and concepts, and explains in detail the operation process of each algorithm and their respective advantages and disadvantages. In current research, researchers' improvements to the algorithm mainly focus on combining more algorithms to complement each other and improving the usability of the algorithm to expand its application scenarios. Overall, after adding machine learning algorithms, the accuracy of vegetable price prediction has been greatly improved, among which the effect of the L-BPNN model is the most prominent, with an average relative error of less than one percent. Such results have considerable reference value for ordinary individual consumers. However, in the face of large-scale production and trading activities, such a level of error may still cause considerable deviation. Current research on vegetable price prediction using machine learning algorithms has a wide range of applications. In subsequent research, introducing new machine learning algorithms should be the main method to improve prediction accuracy. At the application level, improving computing power and improving data collection methods can also improve the actual performance of prediction. In addition, to make this technology benefit more people, developing more efficient and intuitive public information services is also an important improvement plan.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] Chen, J., Li, X., & Liu, X. Research on vegetable price prediction mode based on LSTM and ARIMA. *Journal of Agricultural Engineering*, 2020, 36 (10), 221 - 229.
- [2] Gao, F., Wang, X., & Zhang, X. Vegetable price prediction model based on BP neural network. *Computer and Modernization*, 2020, (11), 89 - 92.
- [3] Zhao, X., Zhang, X., & Wang, X. Vegetable price prediction based on PSO-BP neural network. *Computer and Modernization*, 2018, (12), 95 - 98.
- [4] Wei, G., Huaili, X., Jie, Q. A study on short-term vegetable price prediction based on a combined model of Lasso regression and BP neural network. *Smart Agriculture*, 2020, 2 (3): 108 - 117.

- [5] Ye, L., Li, Y. P. Prediction of vegetable prices based on PSO-BP and RBF neural networks. *Agricultural Network Information*, 2015, 21: 212 - 215.
- [6] Zhang, B., Zhang, L., Fu, Z., & Wang, J. Analysis and prediction of vegetable price change trend based on seasonal index. *Northern Horticulture*, 2017, (18), 185 - 191.
- [7] Cao, X., He, C., & Cui, M. Urban vegetable price fluctuation law and prediction based on X12-ARIMA and LSTM combination model. *Journal of Southwest University for Nationalities*. 2021.