ATLANTIS PRESS

Advances in Computer Science Research, volume 91

International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019)

Raw Grain Price Forecasting with Regression Analysis

Nan Liu[1] and Junwei Yu[2,*]

[1]the PLA Information Engineering University, Zhengzhou, Henan, China

[2]Henan University of Technology, Zhengzhou, Henan, China

*Corresponding author

Keywords: Grain price forecasting, Multivariate linear regression, Neural network, LSTM.

Abstract. Grain price stability and food security are important in all countries. The accurate forecasting of grain price can help the farmer, grain processing enterprise and government make wise decision. A raw grain price dataset is formed with public available data and the raw grain purchase price index is set as the target variable to predict. Three regression models of multivariate linear regression, shallow artificial neural networks and long-short term memory(LSTM) are studied in this paper. Comparative analysis results show that artificial neural network model outperforms the other models in price forecasting on a small dataset. To improve the prediction accuracy of LSTM, the sampling frequency must be increased to get more data to learn the trend and seasonality of grain price.

Introduction

Grains are the harvested seeds of various food crops for human or animal consumption. Generally, raw grain is the collective name of unprocessed food such as wheat, corn, rice soybeans, etc. In China, the annual output of raw grain has exceeded 600 million tons since 2013. Raw grain tends to be the basic natural resource for food products, animal feeding stuff and industrial energy. The grain price is one of the important economic factors that will influence farmer income and company profit. Although almost all the countries have their policies to smooth out the grain price volatility, grain price forecasting analysis is still very important for economic development and food security [1].

Raw grain price forecasting is one of the typical problems of time series analysis. Some models can be used to predict future price given previously observed data. For example, the statistical model of AutoRegressive Integrated Moving Average (ARIMA) is capable of dealing with stationary and univariate time series [2]. As the amount of data is increasing, we can have many variates in time series. Then, we can transform the time series into a supervised learning problem. Many intelligent techniques based on machine learning have been proposed in recent years.

Classification and regression are two major tasks of machine learning[3]. For price forecasting, sometimes we just want to know whether the price will go up or down. It can be treated as a binary classification. While in most situations, we want to predict a continuous and accurate value of the price. So the regression models can be used to predict gran price.

The linear regression can model the relationship between some features and a target response. So we can explore the relationship of grain price and its related factors. Then find the best-fitting line through the sample points. In [3] the linear regression model was used to predict the house price. Artificial neural network (ANN) is one of the most accurate methods for time series forecasting and it has been used in price forecasting of agricultural products [4] and stock [5].

With the development of deep learning [6], the recurrent network and long-short term memory(LSTM) are designed for the processing of sequence data. LSTM is a variant of RNN and provide a solution to the problem of gradient vanishing in RNN. It has proven to be good at emotional analysis and stock price prediction [7][8].

Grain price forecasting is a challenging task. The future price of grain is dependent on many factors such as weather, planting area, supply and demand, and exchange rate. Unfortunately, some important factors are hard to collect from public data. We select the purchase price index, selling price index, exchange rate, and some related consumer price indexes as the features to form a grain

price dataset. The monthly data of grain price are from public web site. As raw grain stands for all of the unprocessed grain, its purchase price is more stationary than the common price of wheat and corn. We focus on the purchase price forecasting of raw grain with different regression models. If we can accurately predict the price of grain, then it is a very good news to increase farmers' income and ensure the profits of grain processing enterprises.

Raw Grain Price and Its Related Factors

The grain purchase price is the actual money paid to grain producers for buying grain with certain weight or volume. In practice, the price index is more common to use. Index of purchase price(PPI) of grain indicates the purchase price relative trend and fluctuation during a given interval of time. In Hebei market price monitoring system of grain and oil[9], PPI is a percentage number that shows the extent to which a price has changed comparing with the price in 2008. 2008 is a reasonable base year for grain price index as we can collect stable and abundant data since that year in China.

The purchase price of raw grain is affected by many factors such as seasonality, supply and demand, policy and macroeconomic situation. Considering the variety of grain circulation, it is hard to obtain accurate data of grain supply and demand. In addition to the historical data, we can take the grain selling price index, consumer price index and exchange rate as the features to predict grain purchase price. Consumer price index(CPI) is one of the most frequently used statistics for identifying periods of inflation. According to the standard of consumer expenditure classification (2013), grain, aquatic, vegetable and fresh fruit belong to the same broad category of food in CPI statistics. National Bureau of Statistics of China [10] publish many detail data about consumer price index every month. These detailed consumer price indexes will be more related to the grain purchase price. So the raw grain price dataset contains the following information: PPI(index of purchase price), SPI(index of selling price), ER(exchange rate), G-CPI(consumer price index for grain),A-CPI(consumer price index for aquatic), V-CPI(consumer price index for vegetables), and F-CPI(consumer price index for fresh fruit).

Grain Price Regression Analysis

In machine learning, regression analysis methods can be used to predict a continuous value of a time series. The past grain purchase price as well as its related features are utilized for future grain price forecasting. We will use some traditional regression methods such as linear regression and neural network. And a deep learning model of LSTM is designed to predict grain price.

Multiple Linear Regression

The relationship between raw grain price and its related factors can be modeled with multiple linear regression. The multiple linear regression model is defined by the following equation:

$$\mathbf{y} = b_1 x_1 + b_2 x_2 + \mathrm{L} + b_k x_k + e = \sum_{i=1}^{k} b_i x_i + e \tag{1}$$

where, b_i represents the coefficient of the explanatory variable. And e is the random error of the prediction.

We can use the classical least-square method to estimate the model parameters of multiple linear regression. The cost function described in (2) is to minimize the sum of squared errors.

$$\mathrm{Q} = \min \sum_{i=1}^{k} (\vec{\mathrm{y}}_i - y_i)^2 \tag{2}$$

where, $\vec{\mathrm{y}}_i$ is the predicted value using formula (1). And y_i is the real price in the dataset. To implement the regression model, we can call function fit of LinearRegression() in the famous machine learning package of scikit-learn.

Artificial Neural Network

Artificial neural networks(ANN) are the biologically inspired models which enables a computer to learn from observational data. ANN has a large number of artificial neurons to process information. A typical ANN consists of input layer, hidden layer and output layer. As traditional ANN does not have so many hidden layers, it is called shallow neural network. ANN can be used to estimate or approximate complex functions that can depend on a large number of inputs. It is one of the most accurate methods to understand relationships between variables, evaluate trends, predict agricultural products price or predict stock price.

Neural networks are good at time series analysis. We can design a price forecasting network with the structure described in Figure 1. Then we can predict the future grain price $y(t)$ from past values of that time series y_i and past values of its related features x_i. The nonlinear regressive function of the network can be written as (3).

$$y(t) = f([x_1, \ldots, x_{t-d}], [y_1, \ldots, y_{t-d}]) \tag{3}$$

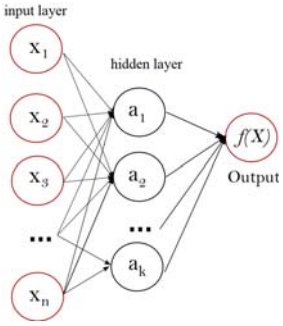

Figure 1. Structure of Artificial neural network.

Here are the steps to design and train a neural network. First, define the structure of network and set the parameters such as the number of neurons in hidden layer. Then, train your network on the training dataset. Calculate the loss function and feed backwards to improve the network. Finally, evaluate the network performance on the test dataset.

Long Short-Term Memory Network

Recently, many models based on deep learning have been proposed for time series[6]. Recurrent Neural Networks(RNN)[11] are a set of powerful artificial neural network algorithms by adding internal feedback loop to their past decisions. There have been incredible success applying RNNs to sequential data processing such as speech recognition, machine translation, natural language processing and video understanding.

Long Short-Term Memory (LSTM) [12] is an improvement of RNN and proves to have good performances in time series learning[13]. As illustrated in Figure 2, LSTM can maintain contextual information as well as temporal behaviors of events. An LSTM layer learns long-term dependencies between time steps in time series. By using the multiplicative gates to improve gradient flow, LSTM layers can learn more context information in a long sequence.

The architecture for a single LSTM cell is shown in Figure 2. An LSTM block typically has one or more memory cells, input gate, output gate, and a forget gate in addition to the hidden state in traditional RNN. In Figure 2, x_t denotes the input; h_t denotes the hidden state which contains the output of the LSTM layer for this time step; c_t denotes the cell state to store the LSTM memory; f is the forget gate; g is the memory cell; i is the input gate and o is the output gate.

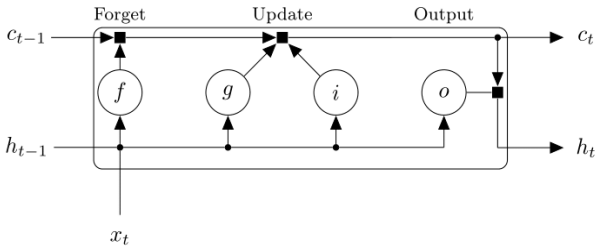

Figure 2. Architecture of LSTM.

At time step t, the LSTM block takes the previous state of the network (c_{t-1}, h_{t-1}) and the current input of the sequence x_t. The output of the layer can add information to or remove information from the cell state by multiplying the weights in the cell state and other gates. Then, update the cell state c_t and h_t. The weights and biases are used to control the extent of information flowing into or remaining in the cell. to compute the output activation of the LSTM block.

Experimental Results and Analysis

First, we will take a look at the grain price dataset, which contains information about raw grain prices index and consumer price index. We collect the data from Grain Information Center of Hebei Provence [9] and the National Bureau of Statistics [10]. The data of price index was sampled every month from 2013 to 2017. The first five lines of the dataset is listed in Table 1. Then we implemented the proposed regression analysis models presented in the previous section on this dataset. The experiments were implemented under Python 3.6 with scikit-learn 0.20.3 and tensorflow 1.13.1 on a computer with Core(TM) i5-4570 and 4GB Ram.

Table 1. The first five lines of the dataset

Date	PPI	SPI	ER	G-CPI	A-CPI	V-CPI	F-CPI
201301	106.61	108.7	627.87	108.8	103.6	104.5	88.3
201302	107.97	109.78	628.45	109.3	104.3	112.8	97
201303	106.46	108.2	627.43	109.5	103.2	105.4	100.2
201304	105.83	106.84	624.71	109.3	102	106	101
201305	105.05	105.77	619.70	109.2	100.8	105.2	102.3

Before implementing different regression models, the data should be preprocessed in advance. First, to handle missing data is very common in data science. The average of weekly data is used for three missing data of exchange rate in 2016. Then, the data are normalized to avoid the influence of the different scale of the original variables. As the value of ER is larger than other features, it will assign more weight to ER without normalization. In order to obtain better fitting and prevent training divergence, we scale the features to a fixed range of [0 1]. The same scaling operation will be applied to the test data. Finally, we split the grain price dataset into a training set and a testing set. The first 80% of time series is considered as training set. And the last 20% is used for model evaluation. That means we use the historical data of the past 4 years to predict the grain price in 2017.

Figure 3 shows the original series of raw grain purchase price index. As we want to predict an exact future value of the price, the loss function for the regression models is defined as mean squared error(MSE). MSE is a useful measurement to calculate the degree of model fitness.

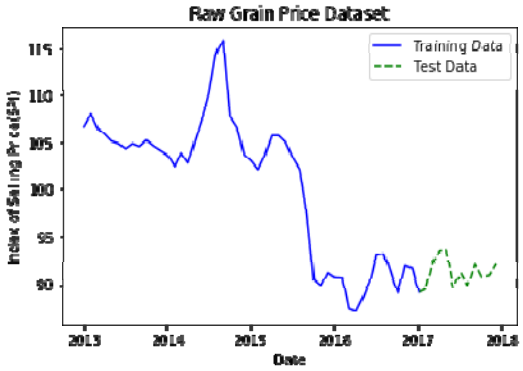

Figure 3. Original series of grain Purchase price

With the linear regression model, the variables are PPI, ER, G-CPI, A-CPI, V-CPI, F-CPI. We fit the multiple linear regression model on the training dataset. The coefficients of the linear regression model are [22.86, 0.77, 1.13, -0.0081, -0.63, 2.22] and the constant term of the model is 87.62. Figure 4 shows the predicted price index and the mean squared error on test data is 1.196.

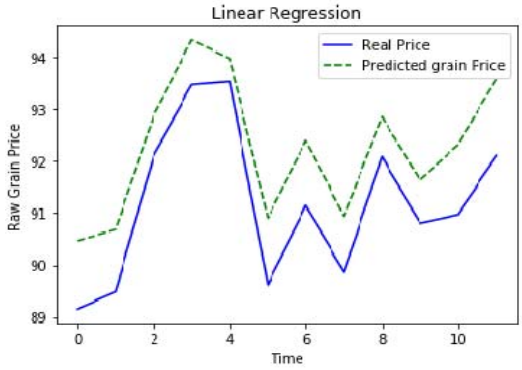

Figure 4. Price forecasting with linear regression model

We design a simple shallow neural network which has only one hidden layer to predict grain price. There are 20 neurons in the hidden layer. The training function is Bayesian regularization backpropagation. Figure 5 shows the predicted price and the mean squared error on test data is 1.095.

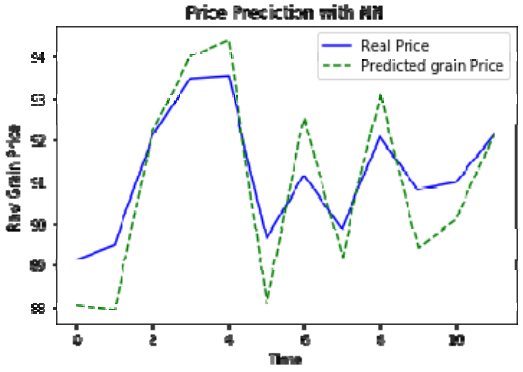

Figure 5. Price forecasting with shallow neural network

As the dataset is small, we define a simple LSTM according to Section 3. The LSTM has four hidden layers which have [20, 20, 50, 20] units to remember the long short memory. The first layer is the input layer to input the 7 features to the network. And the last layer is a regression layer to map a sequence to the output prediction. After training the LSTM, the predicted price index on test data is illustrated in Figure 6 and the mean squared error is 2.821.

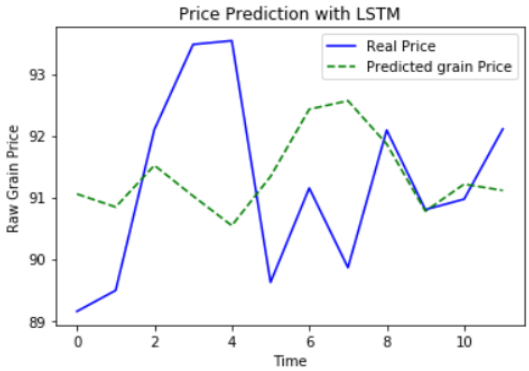

Figure 6. Price forecasting with LSTM

We can calculate the prediction accuracy by the ratio of MSE to the mean purchase price index on test data. The experiment results show that all the prediction accuracies of the proposed models are more than 97%. The multiple linear regression model can learn the trends of the real price, but it shows the obvious positive bias. The model of shallow artificial neural network has a better performance than other models. The LSTM does not show its advantages as it has not enough data to learn the seasonality of grain price and long-term dependence.

Conclusion

Accurate prediction of grain price is essential for avoiding panic selling and unwise purchases. We demonstrate how past publicly-available data can be used to predict future grain prices. Three typical regression models were examined to forecast the purchase price index of raw grain. Experimental results show that all the models are applicable to time series analysis and achieve accurate price prediction results. The results also show that traditional models outperform the deep learning model of LSTM. The primary reason is that there are not enough data for deep learning networks to learn the features and trends of the grain price. In the future work, we will collect the daily or weekly grain prices of more than ten years to train the deep networks. Some advanced models such as temporal convolutional network(TCN) [14]can be used for grain price forecasting. In this paper, we focus on forecasting spot prices of raw grain, and the techniques covered could be applied to more challenging grain futures prices.

Acknowledgment

This research was supported in part by the National Science Foundation under grant 61300123, Cultivation Plan for Young Backbone Teachers of Henan University of Technology-2015. We thank C. Xue and C. Zhao for data preparing and instructive discussion.

References

[1] M. Mallory, Price Analysis: A Fundamental Approach to the Study of Commodity Prices, 2018. http://mindymallory.com/PriceAnalysis/index.html.

[2] G. Box, G. Jenkins. Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco, 1976.

[3] S. Raschka, Python Machine Learning, Packt Publishing, Birmingham, UK, 2015.

[4] G. Li, S. Xu and Z. Li, "Short-term price forecasting for agro-products using artificial neural networks," Agriculture and Agricultural Science Procedia, vol. 1, pp. 278-287, 2010.

[5] G. Qi, L. Zhang, and C. Aggarwal, “Stock price prediction via discovering multi-frequency trading patterns,” 2017. http://www.kdd.org/kdd2017/accepted-papers.

[6] F. Chollet, Deep Learning with Python, Manning Publications, Shelter Island, US, 2018.

[7] Z. Qun, L. Xu and G. Zhang, “LSTM neural network with emotional analysis for prediction of stock price,” Engineering Letters, vol. 25(2), pp. 167-175, 2017.

[8] H. Kim, and C. Won, “Forecasting the volatility of stock price index: a hybrid model integrating LSTM with multiple GARCH-type models,” Expert Systems with Applications, vol. 103, pp. 25-37, 2018.

[9] Information on http://price.heblsj.gov.cn/index/

[10] Information on http://data.stats.gov.cn/easyquery.htm?cn=E0101

[11] W. Zaremba, I. Sutskever, O. Vinyals, “Recurrent neural network regularization,” ICLR, 2015.

[12] S. Hochreiter, and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, pp. 1735-1780, 1997.

[13] S. Srivastava, and S. Lessmann, “A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data,” Solar Energy, vol. 162, pp. 232-247, 2018.

[14] S. Bai, J. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” 2018, arXiv: 1803. 01271.