Frontiers in Business, Economics and Management
ISSN: 2766-824X | Vol. 11, No. 1, 2023

Effectiveness of and bp-neural Network for Overnight Stock Price Forecasting

-- An example of CSI 300

Yaxin Chen*, Yicheng Diao

School of Finance, Southwestern University of Finance and Economics, Chengdu, 611130, China
*Corresponding author: crystal_c2002@outlook.com

Abstract: In view of the empirical problem of introducing machine learning into stock price prediction and the unique condition of Chinese stock market represented by the limitation of rise and fall, In our research, we take CSI300 component stock as samples, using BP neural network to predict stock price, introduce the factor representing the Piece Limit System into the analysis, and demonstrate the effectiveness of Chinese A-share market and the feasibility of neural network to forecast stock price. The study find that the BP neural network further improves the accuracy of stock price prediction with the function of logistic regression, and has a better AUC performance in the sample interval. Our research optimizes the machine learning process to provide more efficient empirical evidence of machine learning in Chinese stock market prediction in recent years.

Keywords: Neural Network, Prediction Model, Big Data.

1. Introduction

Nowadays, with the continuous development of Chinese capital market, stock has become one of the most vital financial instruments for most investors, an important financing method for enterprises, and an essential asset allocation method on the macro level. The ability to accurately predict stock prices has become a hotly researched topic in the stock market.

In empires, much of the discussion on the predictability of stock prices has to do with whether or not the theory of random walk is achieved. In the mid-20th century, a significant portion of scholars believed that stock prices obeyed the random walk assumption and were therefore unpredictable. However, the modern theory is more inclined to hold that stock prices do not obey this assumption and market efficiency is difficult to achieve from the perspective of behavioral finance. Baker, Wurgler, & Yuan(2012) find that people's behavior is often irrational due to various factors and market noise[1].

Before that, other scholars have explored the use of neural network models to discover predictability of stock prices. Kutsurelist (1998) applied a neural network algorithm to predict the S&P index with a 10-day period and obtained good prediction results[2]. Mohamad, Meysam&Ako uses neural network algorithms to predict stock-related variables such as learning per share, expected earnings per share prediction, dividend per share[3]. Fagner, Oliveira&Luis build a neural model for the financial market to predict price behavior, addressing the percentage of correct predictions of price series direction[4].

On the one side, since the daily trading data contains massive noise and nonlinear property, the linear model estimation has deviations cannot be ignored. In that case, scholars use machine learning algorithms to predict stock prices as a supplement to traditional methods, among which, BP neural network is widely applied in stock price prediction. On the other side, Chinese capital market is rife with speculation, which leads to a lot of market anomalies and drastic fluctuations of stock prices. Additionally, compared with the stock markets in other countries, Chinese A-share market has the special trading rule named Price Limit, which sets clear limits on the rise and fall of stocks in every trading day and may affect price prediction.

In view of the problems of current machine learning methods in stock prediction, we optimize the process of machine learning as follows: At first, factor screening is carried out on factors to reduce amount of input variables, improving the running speed of BP neural network, according to study of Buscema, M. (1998)[5] and Li, J., Cheng et al(2012) . Then, the Granger causality and PCA is carried out on the selected influencing factors [7-9]. In the end, the nonlinear relationship between influencing factors and stock price was obtained by Random Forest factor analyze, which have been applied to many studies and achieved better model performance, such as Tanaka et al. (2019) [10]and Beutel et al. (2019)[11]. Additionally, Random Forest also improve the learning speed and prediction accuracy by removing less important factor. Thus, the feasibility of using BP neural network to predict the stock price of Chinese stock market is realized. In view of the influence of the Price Limit Rule and other unique rules on price discovery in the Chinese A-share stock market, in our research we will subjectively add factors representing these systems as a supplement to factors that may affect stock prices.

2. Theory Analysis of BP Neural Network

BP neural network is a network composed of input layer, hidden layer, output layer and weights between nodes of each layer on the basis of Logistic regression. Through forward and back propagation for many times, the relationship between explanatory variables and explained variables approximated to the reality is predicted. The excitation function represented by sigmoid function is nonlinear. Therefore, BP neural network can predict the nonlinear relationship between the

explanatory variable and the explained variable.

In terms of mathematical expression, the values of each node in the forward propagation process and the back propagation process are as follows:

The input value of the i-th node of the hidden layer is expressed as:

$$net_i = \sum_{j=1}^{M} \omega_{ij} x_j + \theta_i \tag{1}$$

The output value of the i-th node of the hidden layer is expressed as:

$$y_i = \phi(\text{ net }_i) = \phi\left(\sum_{j=1}^{M} \omega_{ij} x_j + \theta_i\right) \tag{2}$$

The input value of the i-th node of the output layer is expressed as:

$$y_i = net_k = \sum_{i=1}^{q} \omega_{ki} y_i + a_k = \sum_{i=1}^{q} \omega_{ki} \phi\left(\sum_{j=1}^{M} \omega_{ij} x_j + \theta_i\right) + a_k \tag{3}$$

The output value of the i-th node of the output layer is expressed as:

$$o_k = \psi(net_k) = \psi\left(\sum_{i=1}^{q} \omega_{ki} y_i + a_k\right) = \psi\left(\sum_{i=1}^{q} \omega_{ki} \phi\left(\sum_{j=1}^{M} \omega_{ij} x_j + \theta_i\right) + a_k\right) \tag{4}$$

In these formulas, x_j is the data input on the j-th node of the input layer; ω_{ij} is the weight between the i-th node and the j-th node of the hidden layer; θ_i is the threshold of the i-th node of the hidden layer; $\phi(x)$ is the activation function of each hidden layer; ω_{ki} is the weight between the k-th node of the output layer and the i-th node of the hidden layer; a_k is the threshold of the k-th node of the output layer; $\psi(x)$ is the activation function of the output layer。

In back propagation, correction formula of weight and threshold value is shown as follows:

The weight difference between the k-th node of the output layer and the i-th node of the hidden layer is expressed as:

$$\Delta\omega_{ki} = \eta \sum_{P=1}^{P} \sum_{k=1}^{L} \left(T_k^p - o_k^P\right) \cdot \psi'(net_k) \cdot y_i \tag{5}$$

The threshold difference of k-th node of the output layer is expressed as:

$$\Delta a_k = \eta \sum_{P=1}^{P} \sum_{k=1}^{L} \left(T_k^p - o_k^P\right) \cdot \psi'(net_k) \tag{6}$$

The threshold difference of i-th node of the hidden layer is expressed as:

$$\Delta\theta_i = \eta \sum_{P=1}^{P} \sum_{k=1}^{L} \left(T_k^p - o_k^P\right) \cdot \psi'(\text{ net }_k) \cdot \omega_{ki} \cdot \phi'(\text{ net }_i) \tag{7}$$

In these formulas, η is the learning rate, which controls the speed of neural network fitting.

3. Result

3.1. Data selection and pre-processing

3.1.1. Stock selection

To ensure that sample is representative to China's financial market, the CSI 300 constituent stocks(excluded GEM and KCI stocks) are selected as the stock pool and time range is sat from 17 March 2017 to 31 December 2021. The total number of explanatory variables for each stock initially is 57. For the remaining trading day data, the last 252 trading days data constitutes test set and the rest constructs the training set. All Data come from Wind database.

3.1.2. Factor selection

In order to comprehensively consider the comprehensiveness of the factors affecting the movement of stocks price in next trading day, we have enriched the framework of indicators for predicting the rise(1) and fall(0) of stocks.The initial selection of indicators involves 57 trading indicators,

Table 1. P-value of Granger causality test factor

Factors	p	Factors	p	Factors	p	Factors	p
Pctchange	0.00	Turnover	0.00	Hlimitedays	0.11	Llimitedays	0.60
Pratio	0.00	Avgcost/Close	0.00	Lposition	0.36	Sposition	0.07
Margin	0.02	Tmrate	0.00	Smrate	0.00	Margin_long	0.30
Margin_short	0.50	Net_inflow	0.00	DDX	0.00	DDXR	0.00
Position	0.00	Inflow_rate	0.00	Flowinday	0.00	flowoutday	0.00
Trade_shares	0.07	PE	0.81	PB	0.01	PE_DT	0.09
EV	0.30	BBI	0.43	DDI	0.00	DMA	0.93
MACD	0.00	DMI	0.00	MTM	0.00	BIAS	0.00
CCI	0.00	KDJ	0.00	LWR	0.00	ROC	0.00
RSI	0.00	VRSI	0.04	BRAR	0.00	CR	0.03
PSY	0.00	WAD	0.90	MFI	0.00	MFI	0.00
OBV	0.94	PVT	0.58	WVAD	0.03	BOLL	0.51
CDP	0.64	SAR	0.48	ENV	0.53	MIKE	0.45
ADTM	0.00	DKX	0.54	SKDJ	0.00	Market	0.00
Industry	0.00						

3.2. Factor influence assessment and filtration

3.2.1. Granger causality test

The Granger causality test is a hypothesis-checking statistical method used to test whether one set of economic variables **X** is an influence on another set of economic variables **Y**. In Granger causality test, at 95% confidence interval, 34 explanatory variables out of the original 57 factors have Granger causality with stock rise and fall (p-value < 0.05).

3.2.2. Principal Component Analysis (PCA)

The cumulative contribution of the factors in the PCA test can also be used as one of the indicators for selecting the required factors, and we tend to select the top 85%-95% of the cumulative contribution for use in subsequent models.

Table 2. Factor contribution calculated by principal component analysis

Factor	Factor contribution	Cumulative contribution	Factor	Factor contribution	Cumulative contribution
Pctchange	0.2146	0.2146	DMI	0.0077	0.9414
Turnover	0.1692	0.3837	MTM	0.0076	0.9490
Hlimitedays	0.0652	0.4489	BIAS	0.0074	0.9574
Llimitedays	0.0421	0.4911	CCI	0.0064	0.9628
Pratio	0.0384	0.5294	KDJ	0.0058	0.9686
Avgcost/Close	0.0357	0.5751	LWR	0.0048	0.9733
Lposition	0.0305	0.5957	ROC	0.0045	0.9779
Sposition	0.0295	0.6252	RSI	0.0031	0.9809
Margin	0.0251	0.6503	VRSI	0.0030	0.9839
Tmrate	0.0221	0.6724	BRAR	0.0028	0.9868
Smrate	0.0193	0.6917	CR	0.0026	0.9893
Margin_long	0.0181	0.7099	PSY	0.0023	0.9917
Margin_short	0.0176	0.7274	WAD	0.0020	0.9937
Net_inflow	0.0171	0.7446	VR	0.0014	0.9951
DDX	0.0169	0.7615	MFI	0.0012	0.9963
DDXR	0.0167	0.7782	OBV	0.0010	0.9973
Position	0.0161	0.7943	PVT	0.0008	0.9981
Trade_shares	0.0158	0.8101	WVAD	0.0007	0.9988
Inflow_rate	0.0155	0.8257	BOLL	0.0005	0.9993
Flowinday	0.0144	0.8400	CDP	0.0005	0.9997
Flowoutday	0.0137	0.8536	SAR	0.0002	0.9999
PE	0.0130	0.8666	ENV	0.0000	0.9999
PB	0.0116	0.8782	MIKE	0.0000	1.0000
PE_DT	0.0109	0.8891	ADTM	0.0000	1.0000
EV	0.0102	0.8993	DKX	0.0000	1.0000
BBI	0.0094	0.9088	SKDJ	0.0000	1.0000
DDI	0.0090	0.9177	Market	0.0000	1.0000
DMA	0.0081	0.9258	Industry	0.0000	1.0000
MACD	0.0079	0.9337			

According to PCA result, we removed redundant factors with a cumulative contribution of 90% and retained remained 24 factors. The graph below shows the thermal relationship between the 24 principal components (from left to right, the 0th to 9th principal component respectively) and the 57 factors (from top to bottom, the 0th to 56th factors respectively) derived from the principal component analysis, where blue represents negative correlation (correlation coefficient of 0 to -1) and purple represents positive correlation (correlation coefficient of 0 to 1); the values marked on blocks are the contribution of the factors to the principal components.

3.2.3. Factor importance test using Random Forest model

Random Forest is an integrated learning algorithm using decision trees as the base learner. The model, after training, gives a measure of importance for each feature and features with larger values being more important for prediction accuracy. Noting that the decision tree algorithm is essentially a nonlinear operation, the evaluation of factor contributions using the random forest model allows to filter factors in nonlinearly approach. In this filtration, we select the top 90% of the cumulative importance of factors in reverse order of importance.

3.3. BP neural network model

3.3.1. BP network structure determination

After the grid search method to optimize the parameters, our research finally set a total of 5 hidden layers. Considering the large amount of data in the training set, Random gradient descent (RGD) was chosen as the algorithm used in back propagation to improve the training efficiency of the model, and to compare the advantages of bp neural network compared to simple neural network classification, a logistic regression function was used as the activation function. The learning rate was set to 0.001.

3.3.2. Implicit layer node determination

There is a strong correlation between the number of nodes in the hidden layer of a BP network and the prediction accuracy of the BP network. If the number of nodes is small, the training ability of the neural network is weak and it cannot make accurate predictions with the required accuracy; conversely, if the number of nodes is too large, the training time of the neural network is too long and overfitting may occur.

After a grid search, the final number of nodes per layer of the hidden layer is determined to be 5, 15, 10, 5 and 10 in order.

3.4. Results Analysis

Then we use bp-neural network algorithm to predict whether rise or fall for overnight stocks prices, which we compare with the real change direction. By using the logistic regression function as the activation function for the bp neural network model, the AUC is increased to 0.6128.

4. Conclusions

Our research is dedicated to the study of how various indicators of stocks can be effectively used to make more accurate forecasts of stock gains and losses in order to achieve optimal profitability results. We try to apply BP neural network algorithm to build a model to reflect the more comprehensive characteristics of the stock and to better fit its non-linear fluctuation trend,.

Although our research has achieved some results in using BP neural network to build a model to predict stock prices, showing the relative advantages of the bp neural network classification model over the simple neural network classification model, the model still has shortcomings and do not perform well.

To solve the above problems, Our research suggests that (1) Use less noisy monthly data and include macroeconomic indicators such as inflation rate, GDP growth. (2)For some of the macroeconomic trends or policy influences, dummy variables can be set up to represent them.

References

[1] Baker, M., Wurgler, J., & Yuan, Y. (2012). Global, local, and contagious investor sentiment. Journal of financial economics, 104(2), 272-287.

[2] Kutsurelis, J. E. (2012). Forecasting financial markets using neural networks: An analysis of methods and accuracy. Monterey, California: Naval Postgraduate School; Springfield, Va.: Available from National Technical Information Service.

[3] Ramezanian, M. R., Shaverdi, M., & Faridi, A. (2011). Combination Neural Network and Financial Indices for stock price prediction. Journal of Applied Sciences, 11(19), 3429-3435.

[4] De Oliveira, F. A., Nobre, C. N., & Zárate, L. E. (2013). Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index–Case study of PETR4, Petrobras, Brazil. Expert systems with applications, 40(18), 7596-7606.

[5] Buscema, M. (1998). Back propagation neural networks. Substance use & misuse, 33(2), 233-270.

[6] Li, J., Cheng, J. H., Shi, J. Y., & Huang, F. (2012). Brief introduction of back propagation (BP) neural network algorithm and its improvement. In Advances in Computer Science and Information Engineering: Volume 2 (pp. 553-558). Springer Berlin Heidelberg.

[7] Hiemstra, C., & Jones, J. D. (1994). Testing for linear and nonlinear Granger causality in the stock price-volume relation. The Journal of Finance, 49(5), 1639-1664.

[8] Waqar, M., Dawood, H., Guo, P., Shahnawaz, M. B., & Ghazanfar, M. A. (2017, December). Prediction of stock market by principal component analysis. In 2017 13th International conference on computational intelligence and security (CIS) (pp. 599-602). IEEE.

[9] Wen, Y., Lin, P., & Nie, X. (2020, March). Research of stock price prediction based on PCA-LSTM model. In IOP Conference Series: Materials Science and Engineering (Vol. 790, No. 1, p. 012109). IOP Publishing.

[10] Tanaka, K., Higashide, T., Kinkyo, T., & Hamori, S. (2019). Analyzing industry-level vulnerability by predicting financial bankruptcy. Economic Inquiry, 57(4), 2017-2034.

[11] Beutel, J., List, S., & von Schweinitz, G. (2019). Does machine learning help us predict banking crises?. Journal of Financial Stability, 45, 100693.