Prediction of Hang Seng Index Based on Machine Learning

Weiqin Wang

University of Science & Technology Beijing, Beijing, 100083, China

Keywords: Stock Forecasting; Hang Seng Index**;** Decision Tree; Random Forest

Abstract: The stock market is a complex nonlinear dynamic system. Hong Kong is an open market and Hang Seng Index (HSI) reflects the Hong Kong stock market and the global economy. HSI daily frequency data from January 3, 2005 to July 5, 2019 was used in this paper. Then Decision Tree model and Random Forest model were chosen to predict the movements of HSI based on the highest price, the lowest price, the opening price, the closing price and the volumes. The results showed that Random Forest is more accurate. To some extent, this paper provides insight into investing Hong Kong stock market.

1. Introduction

The stock market is highly chaotic. It cannot ensure its accuracy to make long-term predictions. Short-term predictions, however, are much better. The most concerned issue, the identification of risks, is very important. Now a large number of investment institutions are using various types of model algorithms to predict stock markets. It can be seen that machine learning has great potential. Considering that the stock index includes companies with the largest market capitalization, it is very meaningful to study Hang Seng Index (HSI).

Many predecessors have done a lot of research in this area. For example, Li Yajing used the GARCH model to characterize the volatility of China's stock market. By comparing the differences in the accuracy of each model's volatility prediction, he found out that it can describe China [1-2]. A tool for stock market volatility, but there is a certain time lag between the estimated value of the model and the actual value of the volatility. It is inaccurate to predict the value of the future based on the existing information [3-4]. Yao et al. predicted the follow-up trend of the stock market at a specified time point based on the Markov chain (RICD_MCA), which is based on the rationality index and the characteristic deviation [5]. Wu et.al. predicted the stock market based on the BP neural network [6]. The change of network weight and deviation was calculated in the direction of gradient descent, and the target is gradually approached. This method optimizes the traditional artificial neural network method to predict the stock market trend, but the model should be improved. Wang et al. applied RBF to stock forecasting from the perspective of nonlinear time series. The data in that paper refers to more sTable IBM stock data. Because the stock market is related to various external factors, the method has limited predictive power due to sudden changes in human factors or external economic factors [7-8]. Zhang et al.'s stock forecasting avoids some of the drawbacks of traditional forecasting methods (such as BP). But it also need reasonable analysis and processing of the data [9-10].

Python was used to crawl the HSI daily data of the Hong Kong stock market in this article. After selecting the features, the data was divided into 80% and 20% for training test and testing test, using Decision Tree and Random Forest.

2. Data

HSI daily data of the Hong Kong stock market were selected from January 3, 2005 to July 5, 2019 with a total of 4439 samples. The features involves five variables, including the opening price, the highest price, the closing price, the lowest price and the volume. The data are divided for training set and testing set, 80% and 20% of the total samples respectively. By observing the format of the original data, there is no need to consider the dimension problem and perform normalized

  DOI: 10.25236/aisct.2019.049

operation. As can be seen from the figure, the original data set contains a total of 4439. In order to present the data more clearly, we normalize the data to the maximum and minimum values, as shown in Table 1. The frequency distribution of features can be obtained. And these variables roughly exhibit a normal distribution.

Table 1. Statistical results of stock factors from 2005 to 2019

	Highest	Lowest	Opening	Closing	Volume
count	4439	4439	4439	4439	4439
mean	19980.86	19724.22	19865.35	19856.2	1.43E+09
std	5806.21	5743.66	5780.42	5772.77	3.01E+09
min	8430.62	8331.87	8351.59	8409.01	63873400
25%	14780.77	14527.52	14660.55	14611.42	4.19E+08
50%	21095.9	20853.08	20999.32	20995.01	1.43E+09
75%	23669.46	23409.31	23554.92	23541.15	1.9E+09
max	33484.08	32897.04	33335.48	33154.12	1.92E+11

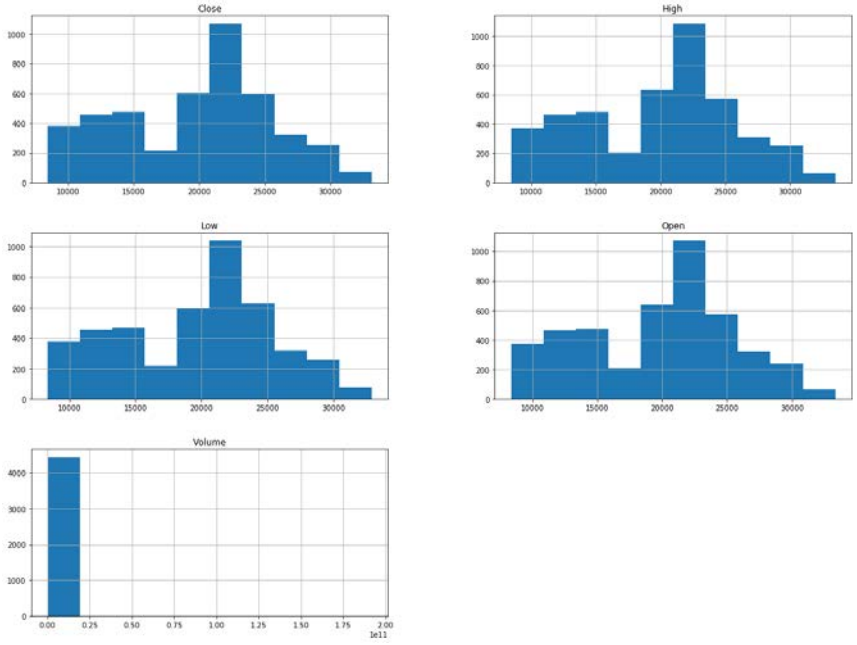

Figure 1. HSI Frequency distribution of features from January 3, 2005 to July 5, 2019

3. Feature selection

Feature selection is used to detect related features, remove extraneous features, obtain feature subsets, and better describe the problem with minimal performance loss [11-13].

3.1 Correlation

We assume that a good feature subset contains features that are highly related to the class, but the features are not related to each other. This is a hypothesis based on heuristics and is at the heart of CFS (correlation-based feature selection) – an inspiring way to assess the value of a subset of features as follows:

$$Merits = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}}.$$

CFS first calculates the feature-class and feature-feature correlation matrix from the training set, and then searches the feature subset space with best first search, which can start with an empty set or a complete set, assuming that the empty set starts, at the beginning No feature selection, and all possible individual features are generated; an estimate of the feature is calculated (represented by

the merit value), and a feature with the largest merit value is selected, and then the second feature with the largest merit value is entered, if the Merit value of these two features is smaller than the original merit value, then the feature of the second largest Merit value is removed, and so on, to find the feature combination that maximizes Merit.

3.2 Linear regression

Linear regression is a regression analysis technique. If the dependent variable of the analysis is a discrete variable, it will be converted into a classification problem. The general form of a hypothesis function is

$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n$.

Loss function is

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

.

Objective function is

$\min J(\theta_0, \theta_1, \cdots, \theta_n)$.

Solving the loss function minimum usually passes the gradient descent method and the normal equation.

Ridge regression and Lasso regression are norm 1 regularization and norm 2 regularization for OLS, respectively. These methods could solve the over-fitting and eliminate the determinant caused by multi-collinearity. The goal can be achieved this by introducing a regularization term in the loss function. The loss function of ridge regression is

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_0(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2$$

.

And the loss function of Lasso regression is

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_0(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j$$

.

3.3 Result

Table 2. Statistical results of Correlation, Linear Regression, Ridge Regression and Lasso Regression

	Linear reg	Ridge	Lasso	Corr.	Mean
Close	**1**	**1**	**1**	**1**	**1**
High	**0.26**	**0.26**	**0.53**	**0.49**	**0.38**
Low	**0.04**	**0.04**	**0.06**	**0.6**	**0.18**
Open	**0.7**	**0.7**	**0.53**	**0.22**	**0.54**
Volume	**0**	**0**	**0**	**0**	**0**

Table 2 showed that the closing price, highest price, lowest price and opening price are important. And these four features were selected and used to predict movements of HSI.

4. Prediction

4.1 Decision Tree

Decision Tree is based on the tree structure for decision making. Generally, a tree contains a root node, several internal nodes, and several leaf nodes; the leaf nodes correspond to the decision results, and each of the other nodes correspond to an attribute test; each node contains a sample set that is divided into child nodes according to the results of the attribute test; the root node contains the sample full set. The path from the root node to each leaf node corresponds to a decision test sequence. The purpose of Decision Tree learning is to create a Decision Tree with strong generalization ability, that is, to deal with the unsuccessful example. The basic process follows a

simple and intuitive divide-and-conquer strategy [14].

4.2 Random Forest

The random “forest” is a machine learning algorithm composed of a number of decision “trees”. It is also an extension of Bagging, that is, multiple Decision Trees are integrated into Random Forests through the principle of Bagging. Among them, Random Forest further introduces random attribute selection in the training process of Decision Tree, which makes Random Forest random, mainly in the two aspects of training sample (random selection sample) and feature selection (random selection feature subset). To ensure that there is no over-fitting and excessive feature dimension. Specifically, the Random Forest uses the bootstrap to extract n samples from the original data, and then uses the Decision Tree to train the extracted n samples. Each Decision Tree determines the final classification by voting. Result. Random Forest has the advantages of simple, easy to implement, and low computational cost. However, for data with more attribute levels, the attribute weight calculated by Random Forest is inaccurate [15].

The mentioned Bootstrap sampling is based on a method of sampling data that is put back, which is the basis of Bagging and Random Forest. Specifically, the n samples are retracted n times to form a new data set. The samples in this new data set may appear multiple times or may not appear. According to statistics, about 63.2% of the samples in the original data set will appear in the new data set. On this basis, the Bagging (Bootstrap Aggregating) algorithm can be constructed, which is to perform multiple Bootstrap sampling on the training sample set. Each time the sampling result can train a weak learner model, thus obtaining multiple independent weak learners (Or weak classifiers, and finally use their combination to make predictions. Among them, when combining the predicted output, Bagging usually uses a simple voting method for the classification task and a simple averaging method for the regression task. If two classes receive the same number of votes in the classification forecast, the simplest method is to randomly select one, and further consider the confidence of the learner vote to determine the final. The basic process of Bagging is as follows:

1) Suppose the sample set, i=1, 2, ..., N, the feature matrix X has H features;

2) Randomly extracting N samples from the sample set with the returned samples to form a sample set Dm;

3) Based on the sample set Dm training model;

4) Repeat steps 2 and 3 to iterate, and average model f(x) to get model F(x)

4.3 Result

Table 3 Evaluation of Decision Tree and Random Forest

	Decision Tree	**Random Forest**
F1	0.65	0.69
accuracy	0.64	0.62
precision	0.65	0.61
recall	0.66	0.79

Under the F1 and recall criteria, the Random Forest works better, as shown in Table 3. The Decision Tree works better under the standards of accuracy and precision. And we use F1-score as the final evaluation method (the average of precision and recall), Random Forest is better than that of Decision Tree.

5. Conclusion

This paper selected the daily data of HSI from Hong Kong stock market from January 3,2005 to July 5, 2019. Then Decision Tree and Random Forest were used to predict the Hong Kong stock index. The result showed that Random Forest is better than that of Decision Tree with F1 as the evaluation index.

However, there is still much room for improvement in this paper. First, we can choose better features and technical indicators, such as MACD, RSI, etc. The MACD indicator is the most used

indicator, and it is the most effective and practical to be tested by history. The indicators are particularly effective in protecting the interests of investors. The RSI reflects the long and short market in terms of price fluctuations over a period of time. The threshold value or its curvature change is considered to determine whether there is an oversold and overbought market, thus implementing its own investment behavior. Therefore, both of them can better help us make predictions about changes in the stock market. Second, we can improve the model. For example, we can choose the XGB model, which is implemented on the basis of Boosting. It can be regarded as the optimization of GDBT, but the difference is that XGB uses the second derivative and regular term, which can be used to control the complexity of the model. Make the learned model easier. We also have the option of CatBoost, which delivers state-of-the-art results and is competitive in performance with other machine learning algorithms with strong accuracy.

References

[1] Yoo P D , Kim M H , Jan T . Financial Forecasting: Advanced Machine Learning Techniques in Stock Market Analysis[J]. IEEE, 2005:1-7.

[2] Refenes A N , Zapranis A , Francis G . Stock performance modeling using neural networks: A comparative study with regression models[J]. Neural Networks, 1994, 7(2):375-388.

[3] Asadi S , Hadavandi E , Mehmanpazir F , et al. Hybridization of evolutionary Levenberg–Marquardt neural networks and data pre-processing for stock market prediction[J]. Knowledge-Based Systems, 2012, 35(none):245---258.

[4] Cao Q , Parry M E , Leggio K B . The three-factor model and artificial neural networks: predicting stock price movement in China[J]. Annals of Operations Research, 2011, 185(1):25-44.

[5] Pan Y , Xiao Z , Wang X , et al. A multiple support vector machine approach to stock index forecasting with mixed frequency sampling[J]. Knowledge-Based Systems, 2017, 122:90-102.

[6] Yu L, Chen H, Wang S, et al. Evolving Least Squares Support Vector Machines for Stock Market Trend Mining[J]. IEEE Transactions on Evolutionary Computation, 2009, 13(1):87-102.

[7] Ince H . Kernel principal component analysis and support vector machines for stock price prediction[J]. A I I E Transactions, 2007.

[8] Hung J C . A Fuzzy Asymmetric GARCH model applied to stock markets[J]. Information Sciences, 2009, 179(22):3930-3943.

[9] Giannetti F , Chirici G , Gobakken T , et al. A new approach with DTM-independent metrics for forest growing stock prediction using UAV photogrammetric data[J]. Remote Sensing of Environment, 2018, 213:195-205.

[10] Chen M Y , Chen B T . A hybrid fuzzy time series model based on granular computing for stock price forecasting[J]. Information Sciences, 2015, 294(2):227-241.

[11] Singh P , Borah B . Forecasting stock index price based on M-factors fuzzy time series and particle swarm optimization[J]. International Journal of Approximate Reasoning, 2014, 55(3):812-833.

[12] Chen Y S , Cheng C H , Tsai W L . Modeling fitting-function-based fuzzy time series patterns for evolving stock index forecasting[J]. Applied Intelligence, 2014, 41(2):327-347.

[13] Li F , Liu C . Application Study of BP Neural Network on Stock Market Prediction[C]// 2009 Ninth International Conference on Hybrid Intelligent Systems. IEEE, 2009.

[14] Wang H , Lu S , Zhao J . Aggregating multiple types of complex data in stock market prediction: A model-independent framework[J]. Knowledge-Based Systems, 2018.

[15] Yixin Z , Zhang J . Stock Data Analysis Based on BP Neural Network[C]// 2010 Second International Conference on Communication Software and Networks. IEEE, 2010.