Predicting the rise and fall of Shanghai composite index based on artificial intelligence

Zijun Wang[1]

[1]Department of International, Hua Qiao University, Quanzhou, 362000, China

Abstract. Shanghai composite index reflects the changes of stock prices, and the methods for various models to predict the stock index emerge one after another, and artificial intelligence is also widely used in various fields due to its stability and accuracy. In this paper, artificial intelligence is applied to Shanghai composite index to predict the stock index. A total of 3422 Shanghai composite indexes from January 1, 2005 to January 1, 2019 were collected, including five indexes: opening price, maximum price, closing price, minimum price and trading volume. Then MA, KDJ and MACD were selected as technical indexes, and their application methods and advantages in Shanghai composite index were analyzed in detail. In addition, in this paper, logistic regression and support vector machine (SVM) in artificial intelligence model were adopted to predict the ups and downs. Finally, it indicates that the support vector basis method based on radial basis is more suitable for stock index prediction model. In this paper, a framework of index prediction is provided by combining technical indicators with artificial intelligence.

1 Introduction

Shanghai composite index, the earliest index published in China, takes all the stocks listed on Shanghai Stock Exchange as the calculation scope, and calculates the composite stock price index [1-3] based on the weight distribution of stock issuance volume. It reflects the overall changes of the stock market, and people can understand the general variation trend of the stock market through Shanghai Composite index. As a result, much attention is paid to the prediction of Shanghai composite index. In recent years, there are a large number of methods to predict the Shanghai composite index, and at the same time, artificial intelligence also gradually infiltrates people's lives. Professor Winston in Massachusetts Institute of Technology evaluates artificial intelligence as "making computers do the research on intelligent work which can only be conducted by human beings in the past". Artificial intelligence involves almost all disciplines of natural science and social science, and its role in the financial field is increasingly prominent. Its main advantages are as follows: first of all, high stability. Artificial intelligence will not be tired, will not be affected by external factors; secondly, it replaces the manual review of financial transactions and related information, so it can better control risks; thirdly, it can more efficiently and accurately analyze and process data. Based on the above advantages, artificial intelligence has been used in intelligent investment, financial market prediction, credit evaluation, convenience services and other financial fields [4-6]. In terms of financial market forecasting, Rebellion, as the world's first purely artificial intelligence-driven fund, once predicted the stock market crash in 2008 and rated Greek bonds an F a month ahead of Fitch. Senoguchi, a machine invented by Mitsubishi, predicted the rise and fall of the Japanese stock market every month, and during the four years, the statistical correctness rate is 68%. Cerebellum, a hedge fund, also used artificial intelligence for forecasting, and it has been profitable since 2009 [7].

At present, there are many main researches on artificial intelligence and index prediction. In 2015, a time series model was established for mutual network search data, and the stock trend was predicted by SVR [8]. Bo Qian et al. investigated the predictability of the Dow Jones Industrial Average index with the use of Hurst exponent for selecting a time period, auto-mutual information. Through these models, they achieved prediction accuracy up to 65 percent [9]. Mu-Yen Chenet al. found that distance partitioning approach neglects the distribution of datasets and can only handle scalar forecasting. They used a novel fuzzy time series model to forecast stock market prices, which is based on the granular computing approach with binning-based partition and entropy-based discretization methods. What's more, their model has been verified by databases in Taiwan Stock Exchange [10]. Ling-Jing Kao et al. came up with a stock price forecasting model in 2013 which integrates wavelet transform, multivariate adaptive regression splines (MARS), and support vector regression (SVR) is proposed to address the problem of wavelet sub-series selection and improve the forecast accuracy. They successfully identify the data of which sessions (or points in time) among past stock market prices exerted significant impact on the construction of the forecasting model [11]. Singh et al. introduced a new Type-2 fuzzy time series model that can

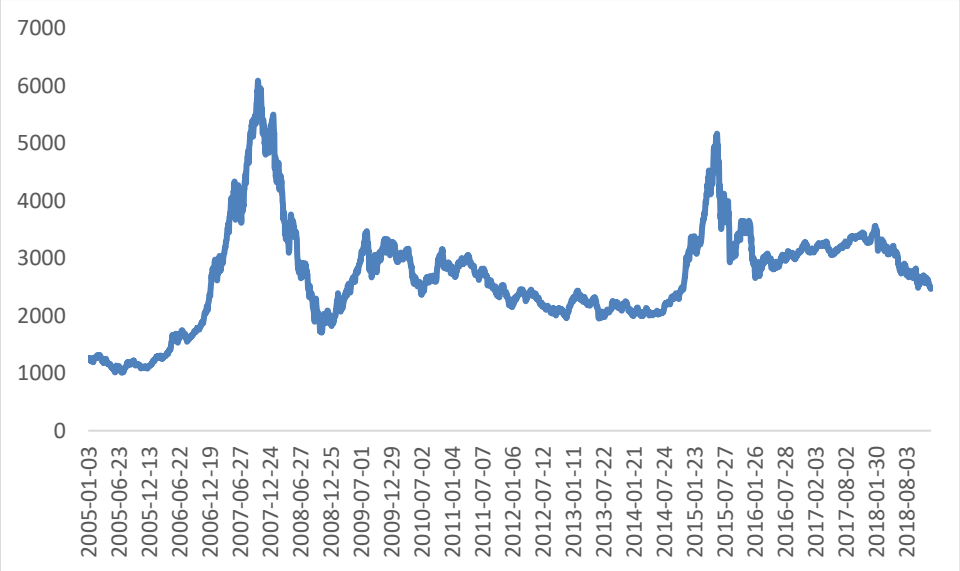
utilize more observations in forecasting and this Type-2 model is enhanced by particle swarm optimization (PSO) technique, which can adjust the lengths of intervals in the universe of discourse that are employed in forecasting, without increasing the number of intervals. Their experimental results demonstrated that the effectiveness and robustness of the proposed model in comparison with existing fuzzy time series models and conventional time series models [12].

Machine learning has been widely applied to index forecasting. Lee et al. used machine learning to predict body mass index status and they found that using logistical regression to build classification models in imbalanced data were effective [13]. Humphries et al. used generalized boosted regression models, a machine-learning algorithm, to calculate a harvest index that takes into account factors that could impact the numbers of birds taken on any given hunt and they successfully got the results. They showed the use of machine learning to correct for extraneous factors (e.g., hunting effort, skill level, or weather) and to create standardized measures could also be applied to other systems such as fisheries or terrestrial resource management [14]. Deo et al. predicted the monthly Effective Drought Index using machine learning algorithms. They concluded that predictions by the extreme learning machine (ELM) was expeditiously efficient and yielded lower prediction errors than the ANN model. Additionally, ELM had faster learning and training speeds than the ANN model [15]. There were also some applications in stock index prediction. Patel et al. compared four prediction models, Artificial Neural Network (ANN), Support Vector Machine (SVM), random forest and naive-Bayes with two approaches for input to these models. The first approach for input data involves computation of ten technical parameters using stock trading data (open, high, low & close prices) while the second approach focuses on representing these technical parameters as trend deterministic data. Experimental results also show that the performance of all the prediction models improve when these technical parameters are represented as trend deterministic data [16]. Araújo et al. proposed a quality index for ornamental rock constructed using machine learning techniques (support vector machines (SVMs)) to model the quality grade allocation procedure as applied by the expert [17].

In this paper, the Shanghai composite index data from 2015 to 2019 was analyzed first, and tMAhen, KDJ and MACD technical indicators were adopted to solve the problem of stock index prediction based on logistic regression and SVM model. In the second part, the application methods and advantages of the three technical indicators were introduced. In the third part, the results of LR and SVM models were compared and analyzed.

2 Data acquisition

The data of Shanghai composite index of 14 years from January 1, 2005 to January 1, 2019 were selected from Sina finance website, covering the opening price, the highest price, the closing price, the lowest price as well as the trading volume. Among the 3,422 pieces of data, there is a little difference in the mean value between opening price and closing price, and the mean value of opening price is about 2,696.82. Among the five groups of indicators, the highest price changed the most, the lowest price has a smallest change, and the variation of opening and closing prices is similar in recent years. In the past 14 years, the highest value of Shanghai composite index is 6,124.04 points and the lowest is 998.23 points. All four groups of indicators are right-skewed, and a considerable part of the data occupies the conceptual data of the first quartile.



Graph 1 Statistics of Shanghai Composite Index from 2005 to 2019.

Table1. Statistical results of stock index from 2005 to 2019.

	High	Low	Open	Close	Volume	Adj Close
count	3422.00	3422.00	3422.00	3422.00	3422.00	3422.00
mean	2722.95	2669.75	2696.82	2699.88	131340.68	2699.88
std	908.17	881.35	896.54	896.76	112224.51	896.76
min	1019.92	998.23	1007.90	1011.50	0.00	1011.50
25%	2139.15	2106.84	2125.65	2127.65	64800.00	2127.65
50%	2743.33	2678.43	2705.49	2707.99	107450.00	2707.99
75%	3182.64	3136.33	3159.72	3161.44	159950.00	3161.44
max	6124.04	6040.71	6057.43	6092.06	857100.00	6092.06

3 Technical index

3.1 MA

Simple moving average, or MA, originally meant moving average, but it is commonly referred to as a moving average because it is usually made into a line. MA is the most basic technical index in the K line diagram, and it is directly drawn on the K line diagram. Compared with K-chart, it can reflect the variation trend of more stable variables (including stock price, trading volume and volume of transaction). The commonly used MA parameters include short-term: MA5, MA10, and Medium and long-term: MA20, MA60.

$$MA\ (N) = (\text{Day one closing price} + \text{Day two closing price} + ... + \text{Day N closing price}) / N$$

By observing the moving average, the market's long/short bias can be studied and judged. Usually, long, medium and short-term moving averages are drawn together. If the three moving averages rise together, it indicates that the market presents a bullish pattern. On the contrary, if the three moving averages fall side by side, the market shows a pattern of short positions. The moving average can recognize the end or reversal of the old market trend as well as the emergence of a new one. It cannot be ahead of the market, but it can be a faithful reflection of the market. Because the moving average is a smooth curve, it can effectively filter out obscure small jumps and objectively reflect the market's general trends.

3.2 KDJ

Random indicator KDJ is commonly used in the statistical system of stock analysis. According to the principle of statistics, through the proportional relationship among the highest price and the lowest price appeared in a special period (usually 9 days, 9 weeks, etc.) and the closing price of the last calculation cycle, the immature random value RSV of the last calculation cycle is calculated. Then according to the smooth moving average method, K value, D and J values are calculated, and the graphs are drawn to study and judge the trend of stocks.

To calculate KDJ, firstly, the RSV value of the period, namely the immature random index value, shall be calculated, and then K value, D value and J value shall be calculated in turn. Taking the calculation of daily line data of KDJ for example, the calculation formula is as follows:

$$RSV = \frac{\text{the closing price on that day} - \text{the lowest price}}{\text{the highest price on the recent N days} - \text{the highest price on the rencent N days}} \times 100$$

$$K \text{ value on the day} = 2/3 \times K \text{ value on the previous day} + 1/3 \times RSV \text{ on the day}$$

$$D \text{ value on the day} = 2/3 \times D \text{ value on the previous day} + 1/3 \times K \text{ value on the day}$$

$$J = 3D - 2K$$

3.3 MACD

MACD is called the moving average convergence divergence, which developed from the double exponential moving average. The meaning of MACD is basically the same as that of the double moving average, that is, the dispersion and aggregation of the fast and slow moving average are used to represent the current long-short state and the possible development trend of stock price. The changes in MACD represent the changes in market trends, and MACD at K-line level represents buying and selling trends in the current level cycle. MACD is the DIF of the average market cost and generally reflects the overall trend of the medium-term stocks. The advantage of combining MACD with KDJ to judge the market is that it can more accurately grasp the short-term buying and selling signals of KDJ index. At the same time, due to the medium-term trend reflected by the characteristics of MACD index, the two indexes can be used to determine the medium and short-term fluctuations of stock prices. The calculation formula is as follows:

$$EMA\ (N) = \alpha\ (C - EMA') + EMA'$$

$$\alpha = \frac{2}{N+1}$$

$$\mathrm{DIF} = \mathrm{EMA}（12）- \mathrm{EMA}（26）$$

$$\mathrm{DEA}（N）= \frac{2}{N+1} DIF + \frac{N-1}{N+1} DEA'$$

$$BAR（MACD）= 2 * (\mathrm{DIF} - \mathrm{DEA})$$

Where EMA index moving average is also called EXPMA index; α is the smoothing coefficient; C is today's closing price' EMA' represents yesterday's EMA; and EMA(N) represents EMA on day N.

3.4 Results

In this paper, 3412 MA, KDJ and MACD data were collected. Among them, the mean value of MA is 0.41, the mean value of other indexes is higher than that of MA, and the standard deviation of MA is 3122.12, with the largest variation range. The standard deviation of KDJ and MACD is 0.4 and 0.6.

Table2. The statistical results of technical indicators in each day.

	COUNT	MEAN	STD
MA	3412	3122.12	123.49
KDJ	3412	1.44	0.4
MACD	3412	0.5	0.6

4 Artificial intelligence model

4.1 Logistic regression

Logistic regression model is often used to deal with classification problems, such as predicting whether stocks will rise or fall in the next month, so it is a machine learning method which is often used to solve classification problems. In linear regression, 0.5 is used as the threshold value to judge positive and negative examples, but 0.5, as the threshold, may lead to wrong sample classification, while logistic regression can compress the prediction range from real number field to the range from 0 to 1, so as to further improve the predicted quasi-curvature. The formula is as follows:

$$y = \frac{1}{1 + e^{-(\theta^T x + b)}}$$

Where Y is the decision value; x is the eigenvalue; e is the natural logarithm; w is the weight of the eigenvalue; and b is the bias. $\theta（x）$ is the inner product of the two.

The advantages are as follows: first of all, the possibility of classification can be modeled directly, so there is no need to realize the distribution of assumed data, and the problem caused by inaccurate distribution of assumed data is also avoided. Secondly, the form is simple, the model is highly interpretable, and the influence of different features on the final results can be seen from the weight of features. Moreover, in addition to categories, approximate probability prediction can be obtained, which is beneficial to many tasks which need to use probability to assist decision-making.

4.2 Support vector machine

Support vector machine (SVM) is a binary classification model, which aims to find a hyperplane to segment the sample. The principle of segmentation is to maximize the interval, and finally transform it into a convex quadratic programming problem to solve. The idea of support vector machine is to classify hyperplanes and introduce more dimensions, and then introduce kernel methods to calculate hyperplanes.

First of all, hyperplane is divided in the sample space: $W^T x + b = 0$, where W is the normal vector, which determines the direction of hyperplane, and b is the displacement, which determines the distance between hyperplane and origin. It is assumed that hyperplane can correctly classify the training samples, that is, for training samples, the following formula is satisfied:

$$\begin{cases} W^T x_i + b \geq +1 & y_i = +1 \\ W^T x_i + b \leq -1 & y_i = -1 \end{cases}$$

The following formula is deduced:

$$\begin{cases} w^T x_+ = 1 - b \\ w^T x_- = -1 - b \end{cases}$$

Then the interval is obtained as follows:

$$\gamma = \frac{1 - b + (1 + b)}{\| w \|} = \frac{2}{\| w \|}$$

In order to maximize the interval, the basic model of support vector machine is obtained as follows:

$$\min_{w,b} \frac{1}{2} \| w \|, s.t. y_i (w^T x_i + b) \geq 1 (i = 1,2,..., m)$$

The kernel function is as follows:

$$\kappa(x_i, x_j) = < \phi(x_i), \phi(x_j) > = \phi(x_i)^T \phi(x_j)$$

The SVM doesn't make any assumptions about the distribution of the original data, which indicates that the SVM model has a low requirement for data distribution and a wider applicability. If there is no prior information about the data distribution in advance and the distribution is unknown, then the SVM model can be used for data processing. And SVM is very efficient in high-dimensional space.

5 Results

As it is based on logistic regression model, there are many common parameters. Penalty term C is the penalty coefficient used to control the loss function, and it is similar to the regularization coefficient in LR. If C value is large, it is equivalent to the penalty slack variable, which is expected to be close to 0, that is, the penalty for misclassification is increased, which tends to be the case of complete correct classification of training sets. In this way, the accuracy of training set testing is very high, but the generalization ability is weak, which easily leads to overfitting. If C value is small, the penalty for

misclassification is reduced, the fault-tolerant ability is enhanced, and the generalization ability is strong, but it may also be underfitting. Among all the models in this paper, C is selected as 0.1. There are three main methods for support vector classification in sklearn: SVC, NuSVC and LinearSVC, which are expanded as three support vector regression methods: SVR, NuSVR and LinearSVR. The results based on logistic regression and SVM are shown in Table 2.

Table3. Comparison of SVM and RF results.

	LR	SVM linear	SVM rbf
Accuracy	0.77	0.81	0.92
Precision	0.75	0.82	0.93
Recall	0.76	0.83	0.89
F1 Score	0.75	0.84	0.92

As can be seen from Table 2, SVM is better than SVM linear in terms of accuracy, precision and recall, and the performance of LR in all aspects is lower than that of SVM. Among them, the F1 Score of SVM based on RBF is up to 0.92, which is higher than SVM based on linear (0.84) and LR (0.75). Therefore, the SVM model based on RBF is more suitable for stock prediction.

6 Conclusions

In this paper, a total of 2,433 Shanghai Composite indexes from 2005 to 2019 were selected, and MA, KDJ and MACD were selected as the technical indexes, and then the two methods: logistic regression and support vector machine in the artificial intelligence model were analyzed. It was found that support vector machine was more suitable for stock index prediction and it was used as the basis for Shanghai composite index prediction.

In this paper, by combining artificial intelligence with finance, the machine learning method was adopted to predict the stock index, and the efficiency of support vector machine (SVM) was made full use of, but the results may not be accurate enough, and some new models can further improve the results, such as decision tree. Through cross validation, the accuracy was improved, and random forests avoided overfitting. Meanwhile, it can consider the importance of variables and the operation is relatively simple. xgboosting model considers the situations when training data is sparse values, and it can specify the missing value or the value of the specified default direction of the branch, which can greatly improve the algorithm's efficiency and avoid overfitting phenomenon; In addition, new indicators, such as PSY, emotional indicator of investors' psychological fluctuations in the stock market and other variables, can also be used to describe stocks, and predict the index under a more accurate framework.

References

1. Fethi M D, Pasiouras F. Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey[J]. European journal of operational research, 2010, 204(2): 189-198.
2. Applications of artificial intelligence in finance and economics[M]. Emerald Group Publishing Limited, 2004.
3. Krollner B, Vanstone B J, Finnie G R. Financial time series forecasting with machine learning techniques: a survey[C]//ESANN. 2010.
4. Trippi R R, Turban E. Neural networks in finance and investing: Using artificial intelligence to improve real world performance[M]. McGraw-Hill, Inc., 1992.
5. Bahrammirzaee A. A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems[J]. Neural Computing and Applications, 2010, 19(8): 1165-1195.
6. Trippi R R, By-Lee P, Jae K. Artificial intelligence in finance and investing: state-of-the-art technologies for securities selection and portfolio management[M]. McGraw-Hill, Inc., 1995.
7. Kim K. Artificial neural networks with evolutionary instance selection for financial forecasting[J]. Expert Systems with Applications, 2006, 30(3): 519-526.
8. Liu Y , Chen Y , Wu S , et al. Composite leading search index: a preprocessing method of internet search data for stock trends prediction[J]. Annals of Operations Research, 2015, 234(1):77-94.
9. Qian B , Rasheed K . Stock market prediction with multiple classifiers[J]. Applied Intelligence: The International Journal of Artificial, Intelligence, Neural Networks, and Complex Problem-Solving Technologies, 2007, 26(1):25-33.
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. Kao L J , Chiu C C , Lu C J , et al. A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting[J]. Decision Support Systems, 2013, 54(3):1228-1244.
12. 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.
13. Lee B J , Kim K H , Ku B , et al. Prediction of body mass index status from voice signals based on machine learning for automated medical applications[J]. Artificial Intelligence in Medicine, 2013, 58(1):51-61.
14. Humphries G R W , Bragg C , Overton J , et al. Pattern recognition in long-term Sooty Shearwater data: applying machine learning to create a harvest index[J]. Ecological Applications, 2014, 24(8):2107-2121.
15. Deo R C , Şahin, Mehmet. Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia[J]. Atmospheric Research, 2015, 153:512-525.
16. Patel J , Shah S , Thakkar P , et al. Predicting stock

and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques[J]. Expert Systems with Applications, 2015, 42(1):259-268.

17. Araújo, M, Matías, J. M, Rivas T , et al. Machine learning techniques applied to the construction of a new geomechanical quality index[J]. International Journal of Computer Mathematics, 2011, 88(9):1830-1838.