2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS) | 978-1-7281-9085-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICICAS51530.2020.00019

# Research and prediction of Shanghai-Shenzhen 20 Index Based on the Support Vector Machine Model and Gradient Boosting Regression Tree

Pingwen Xue, Yuan Lei*, Yanxing Li
School of Information and Engineering,
Guangxi University of Foreign Languages,
Naning,Guangxi,530007,China
Corresponding Author* leiyuandata@foxmail.com

***Abstract*—The non-stationarity, non-linearity and multi-factor influence of financial index series makes it more difficult to predict. In order to achieve a more accurate stock index prediction, this paper selects a public data set provided by a well-known securities company to predict the CSI 20 Index (CSI20) based on the gradient-enhancing regression tree model and support vector machine model. RMSE, MAE, and MAPE are selected as the indicators to evaluate the differences in the forecasting ability of the two models in the financial sector. The results of this study can help investors to adopt effective investment strategies and reduce investment risks.**

***Index Terms*—stock index forecasting, machine learning, gradient-improvement regression tree, support vector machine**

## I. Introduction

The stock market is an important and indispensable player in China's financial market. How to effectively manage risk in the stock market has always been a key concern for financial regulatory authorities and various investment entities. Both investors and speculators are present and highly active in the stock market. The many players in the stock market follow different stock trading philosophies in investing and playing with each other, resulting in perennial volatility in the stock market. These fluctuations are not only opportunities, but also risks in the stock market [1]. As a barometer of the national economy, the stock market is becoming more and more important in the social economy. Fluctuations in the stock market can have an impact far beyond people's expectations and may even become a destabilizing factor in society.

A stock index is an average stock price index calculated by weighting some representative stocks among many stocks as constituents according to certain rules. In this paper, the CSI20 index is proposed by a Chinese securities firm to reflect the movement of high-quality stocks in the Internet and pharmaceutical sectors on the Shanghai Stock Exchange. With the improvement of people's living standards, the Internet and pharmaceutical industries have become popular investments today. The index can reflect the overall trend of the two industries, provide investors with a reference for investing in individual stocks, and achieve risk diversification and investment transparency.

Although some researchers have argued that stock market forecasting is unattainable according to the Efficient Markets Hypothesis (EMH) [2]. However, according to the results of behavioral finance and computational experimental finance empirical studies, financial markets have a certain probability of predictability [3], [4]. With the development of stock market research, the field has now evolved to be dynamic, nonlinear, complex, and inherently chaotic. Many studies have begun to focus on the use of traditional or emerging techniques to improve the quality of stock index predictions [5]. Traditional measurement models such as linear regression, autoregression, etc. can be used to test models and parameters based on classical statistical principles. However, traditional methods need to take into account many immeasurable factors, such as financial time series, smoothness of data, etc. This makes traditional models useful in real life. This makes the application of traditional models in real life very limited [6]. And with the deepening of related research, it is found that the traditional measurement model does not solve the nonlinear problem well, so the traditional model has great limitations in the stock index prediction problem.

In this paper, to address the shortcomings of traditional models, artificial intelligence techniques that do not need to meet statistical assumptions and can better handle nonlinear problems are beginning to shine. Artificial intelligence methods are based on machine learning to train and model historical data, and can provide better accuracy for nonlinear time series data. Typical models are Random Forests, Gradient Boosted Regression Trees (GBRT), Artificial Neural Network (ANN), Support Vector Machine (SVM), etc. [7], [8]. As the research of AI technology progresses, it is found that many AI techniques are still inadequate for high latitude data and for the immunity of anomalous data [9]. Finally, the gradient-raising regression tree and the support vector machine can avoid these problems.

978-1-7281-9085-3/20/$31.00 ©2020 IEEE
DOI 10.1109/ICICAS51530.2020.00019

## II. Related concepts and theoretical basis

### A. Support vector machine

According to the calculation rules of stock index mentioned above, the calculation of stock index is a regression problem. However, after literature review, we find that there are many factors that need to be considered in the calculation of stock index, so the problem of stock index calculation is actually a nonlinear regression problem. The following will introduce the theory of support vector machine model under nonlinear conditions.

For nonlinear problems, support vector machine (SVM) maps the sample space into the high-dimensional feature space (Hilbert space) by selecting appropriate nonlinear mapping, and carries out linear regression in this space. Thus, the nonlinear regression problem in the low dimensional space is transformed into the linear regression problem in the high dimensional space. Support vector machine (SVM) is as follows.

$$\begin{array}{l} \min_{v,b,\zeta} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \\ st \quad y_i\left(w\cdot\phi\left(x_i\right)+b\right) \geq 1-\xi_i \quad \xi_i \geq 0 \quad i=1,2,\ldots,N \end{array}$$

According to the calculation, we can get the classification decision function in the following form.

$$f(x) = \operatorname{sign}\left(\sum_{i=1}^{N}\alpha_i^* y_i\left(\phi(x)\cdot\phi\left(x_i\right)\right)+b^*\right)$$

The method is used to make the classification decision, and the inner product of the calculated sample points is placed in the high space. If the dimensions in the space are very high, the calculation amount of the model will become very large, and there is a great probability of "dimension disaster". In order to avoid this problem, the kernel function is introduced into support vector machine to reduce the calculation amount of the model.

$$K\left(x_i, x_j\right) = \phi\left(x_i\right)\cdot\phi\left(x_j\right)$$

The common kernel functions include polynomial kernel function, Gaussian radial basis function (RBF) kernel function and sigmoid kernel function.

$$K(x,z) = (< x,z > +R)^d$$

$$K(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$$

$$\mathrm{K}(x,z) = \tanh(v(x,z)+c)$$

### B. Gradient Boosting Regression Tree (GBRT)

Gradient lifting regression tree is an iterative regression tree algorithm. It was considered as a generalization algorithm as SVM at the beginning. The algorithm is composed of multiple regression trees, and the conclusion of all trees is accumulated to get the final result. Gbrt is an algorithm in boosting family.

Boosting algorithm is a multi-classification combination algorithm, in which a single classifier is called "weak classifier", and the classifier combined by multiple weak classifiers is called "strong classifier". The weak learning algorithm is promoted to the strong learning algorithm. However, the algorithm needs to assume the lower limit of error in advance, and its application in real life is limited.

After that, people improved the algorithm and proposed AdaBoost algorithm, which can adjust the feedback according to the results of weak learning. No prior knowledge of weak learner performance is required. The lifting tree used in this paper is based on the algorithm model.

Binary regression tree is usually used in the application of the lifting tree model to regression problems, and the corresponding name of the lifting tree model is lifting the regression tree.It can be expressed mathematically as an additive model of a decision tree $\mathrm{f}_M(x) = \sum_{m=1}^{M} T\left(x,\Theta_m\right)$, $T\left(x,\Theta_m\right)$ where it represents a decision tree.$\Theta_m$is the parameter of the decision tree, M is the number of decision trees.

For weak classifiers, the additive model and forward distribution algorithm is combined in the lifting tree, and the initialization function can be assumed to be, then the iterative function of $F_0(x)$.Step M is as follows.

$$F_m(x) = F_{m-1}(x) + h\left(x,\theta_m\right)$$

The algorithm uses a loss function to optimize the next learner parameter, which is as follows.

$$\theta_{m+1} = \operatorname{argmin}\sum_{t=1}^{N} L\left(y_t, F_m(x) + h\left(x,\theta_{m+1}\right)\right)$$

After M iterations, the decision tree is integrated to get the regression problem lifting tree $\mathrm{f}_M(x) = \sum_{m=1}^{M} T\left(x;\Theta_m\right)$.

In gradient lifting regression model, the choice of loss function is also a crucial step. Since the stock index prediction problem in this paper is a regression problem, we use the square error as the loss function in this model. $L(y,f(x)) = (y-f(x))^2$.The model is modified for the next model through calculation, that is $\mathrm{r} = y - f_{m-1}(x)$,the residual error of the current model fitting.Generally speaking, a good model is what we want the model to get $T\left(x,\Theta_m\right) \longrightarrow r$.When the loss calculation approaches zero, this is the result of model optimization.

## III. Empirical analysis of Shanghai and Shenzhen 20 index

The data used in this paper is from the second Guangxi University Students artificial intelligence competition. The

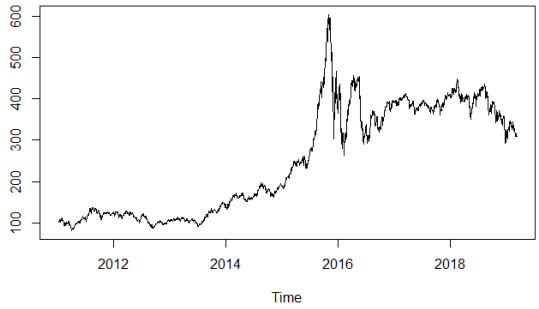


Fig. 1. CSI 20 chart

data set is provided by a well-known securities company and includes financial data from April 2011 to December 2019.In the empirical process, the data from 2011 to 2018 in the data set are used as the training set, and the data in 2019 as the test set. Train the model on the training set, and use the model to predict the trend of csi20 in 2019.The predicted value is compared with the real value, and the prediction of the model is analyzed and evaluated. Because the data dimension of the given data set is too high, in order to prevent the model from invalid due to the dimension explosion and gradient disappearance, this paper uses CSI300 rule to clean the data set and screen out the redundant data.

### *A. descriptive statistics of data*

The trend of csi20 index in the selected period is shown in the figure 1.

From the chart figure 2, we can see that in the selected time period, csi20 index data has experienced three stages of rise, fall and rapid rise, and the data demonstrate a great fluctuation.

```
> summary(zs$hushen20)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  82.07  119.69  209.79  248.89  382.35  602.84
```

Fig. 2. CSI 20 stock index data analysis

From the minimum and maximum, we can find that the exponential data span is very large. From the perspective of standard deviation, we can find that the stock index fluctuates violently. In order to view the data more intuitively, we visualize the stock index data, as shown in the figure 3.

It is clear from the figure that the data does not obey a normal distribution, so at any level of significance, it is possible to make a decision to reject the data as obeying a normal distribution, and therefore use historical data to make predictions about the future.

### *B. Gradient lifting regression tree*

According to the gradient elevation tree, we can get the order of importance of variables(figure 4), and we can see that X1-X20 is arranged from top to bottom in the order of smallest to largest, which basically meets the rules of Chinese stock market, so no further data filtering is needed.

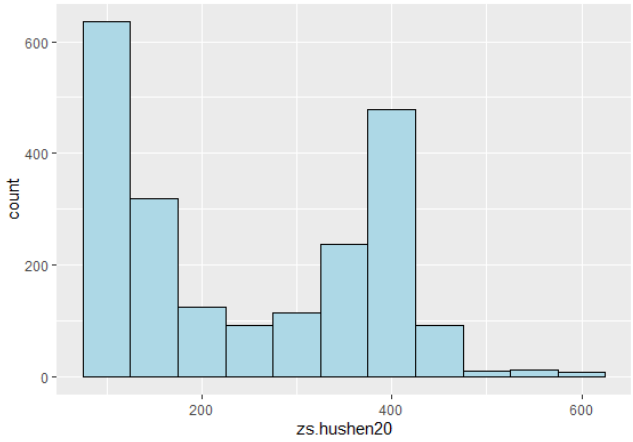


Fig. 3. CSI 20 stock index data distribution map

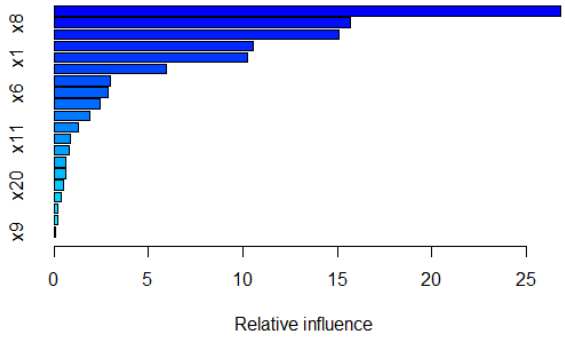


Fig. 4. Importance ranking

In addition, we use the method of ”min-max” to model the regression tree, which is the most suitable for the regression tree (figure5;figure6).

The modeling results are shown in Figure 7. The black line shows the real data of the CSI20 stock index, and the red line shows the model’s prediction of the CSI20 data.

As can be seen from the figure, although the trend of stock index data is not well predicted, the range of data gap is relatively small. The error is less than 50, which shows that the model can play a certain reference role in stock index prediction.

### *C. Support vector machine*

Using the improved RBF kernel function support vector machine model and CSI20 data training set, the model parameters are shown in Figure 8. In the support vector machine model parameters, we can see that the ”gamma” value of 0.05 is very small, which shows that the SVM model established in this paper is more reasonable and can be used for stock index prediction.

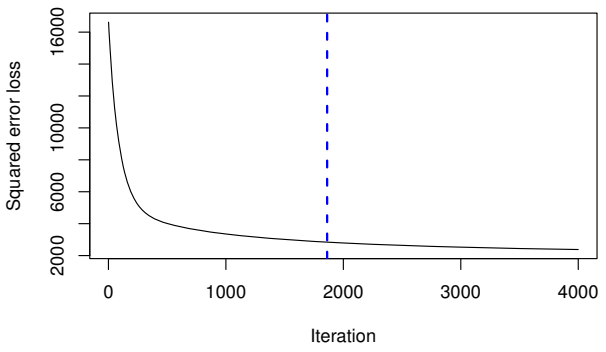


Fig. 5. Gradient lifting regression tree value graph

```
> print(best.iter)
[1] 1862
attr(,"smoother")
Call:
loess(formula = object$oobag.improve ~ x, enp.target = min(max(4,
    length(x)/10), 50))

Number of Observations: 4000
Equivalent Number of Parameters: 39.99
Residual Standard Error: 1.117
```

Fig. 6. Specific parameter values of gradient lifting regression tree

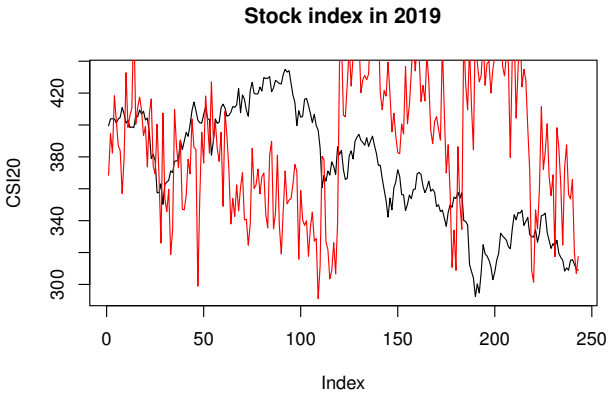


Fig. 7. Gradient lifting regression tree prediction results

In this paper, the value of radial basis function in Gaussian kernel function only depends on the distance from the origin. In its expression, $\|\mathrm{x}-\mathrm{z}\|^2$ it refers to the square Euclidean distance between the observed data X and Z. The if x and the Z closes$\|x-z\| \approx 0$.Then the kernel value is 1; if x and Z are quite different$\|\mathrm{x}-\mathrm{z}\| \gg 0$.Therefore, the radial basis function of the financial index in this paper can theoretically map the original features to infinite dimensions. As shown in the figure 9.

The rest of the radial basis functions can be used according to the data usage. The program that can complete these tasks

```
> summary(svr)

Call:
svm(formula = V2 ~ ., data = dat1.5)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.05
    epsilon:  0.1


Number of Support Vectors:  1178
```

Fig. 8. SVM model modeling results

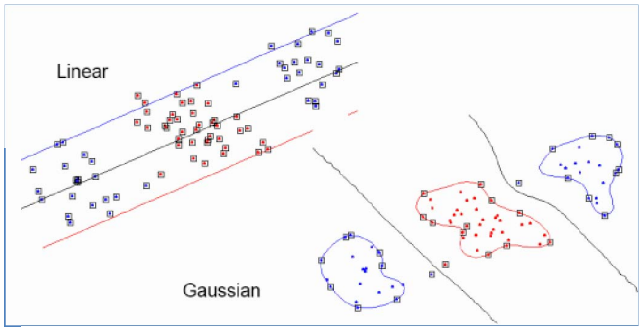


Fig. 9. Gaussian kernel function

well is called support vector machine (graph).

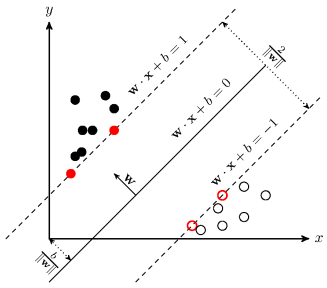


Fig. 10. Support vector machine implementation

Therefore, the dual optimization problem of financial index support vector machine is as follows.

$$\begin{gathered}\min \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K\left(x_i \cdot x_j\right)-\sum_{i=1}^{N} \alpha_i \\ \text{s.t.} \quad \sum_{\substack{j=1 \\ \alpha_i \geq 0}}^{N} \alpha_i y_i=0 \\ i=1,2, \ldots, N\end{gathered}$$

The introduction of kernel function effectively reduces the computational complexity of the model in low dimensional space. The model classification display is carried out in high latitude space, and the calculation amount is not affected by the dimension space, so the dimension explosion is effectively avoided.

TABLE I
ERROR COMPARISON

| Error calculation | Gradient lifting regression tree | Support vector machine regression |
|---|---|---|
| RMSE | 64.3918 | 23.9340 |
| MAE | 54.0054 | 18.5488 |
| MAPE | 14.9435 | 4.9115 |

In this paper, the model is also used to predict the data of csi20 stock index in 2019, and the model fitting data are shown in figure 11 (true value of black line, predicted value of red line).

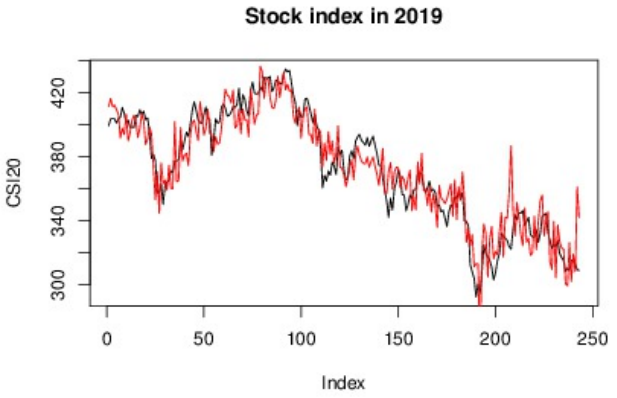


Fig. 11. Support vector machine prediction results

It can be seen from the figure that the predicted value of SVM model is basically consistent with the curve drawn by the actual value. The support vector machine can predict the stock index effectively.

### D. Model comparison

In this paper, RMSE, Mae and MAPE are used as the evaluation criteria of the model

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

In this paper, multi criteria are used to evaluate the model, and the results are more comprehensive and more reasonable than single comparison.

From table I, we can see that the performance of support vector machine is better than that of the three evaluation criteria selected in this paper. Compared with gradient lifting regression tree, support vector machine is more suitable for high latitude data regression. In addition, only a few support vector machines need the same amount of support vector computation. These characteristics of support vector machine make it can avoid "dimension disaster" in most cases, and the problem of "gradient disappearance" of regression tree for the gradient promotion still needs to be paid attention to. In this paper, we consider that the performance of support vector machine is better than that of gradient lifting regression tree.

## IV. CONCLUSION

This paper discusses the performance of support vector machines and gradient boosting regression trees in dealing with stock index forecasting problems. The CSI20, which is representative of the Internet and pharmaceutical sectors that are currently hot in the stock market, is selected for forecasting. The relationship between the constituent stocks and the stock indices is modeled. The regression model based on the RBF kernel function model with support vector machine model and the gradient upgrading regression tree model has good predictive ability. The results were validated using several evaluation criteria. Future research should focus on exploring other algorithms and models for solving stock market forecasting problems.

## V. ACKNOWLEDGEMENTS

Young and middle-aged teachers' basic ability improvement project of Guangxi colleges in 2020(2020KY63021).

## REFERENCES

[1] S. Iamsiraroj, "The foreign direct investment–economic growth nexus," *International Review of Economics & Finance*, vol. 42, pp. 116–133, 2016.
[2] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.
[3] J. J. Murphy, *Technical analysis of the financial markets: A comprehensive guide to trading methods and applications*. Penguin, 1999.
[4] J. Felsen, "Learning pattern recognition techniques applied to stock market forecasting," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-5, no. 6, pp. 583–594, 1975.
[5] B. Yang, W. Zhang, and H. Wang, "Stock market forecasting using restricted gene expression programming," *Computational Intelligence and Neuroscience*, vol. 2019, 2019.
[6] G. CHI and Z. LI, "Forecast model of stock index futures prices based on small sample," *ICIC express letters. Part B, Applications : an international journal of research and surveys*, vol. 5, no. 3, pp. 657–662, jun 2014. [Online]. Available: https://ci.nii.ac.jp/naid/40019968889/en/
[7] J. Cao and J. Wang, "Stock price forecasting model based on modified convolution neural network and financial time series analysis," *International Journal of Communication Systems*, vol. 32, no. 12, p. e3987, 2019.
[8] M. Mallikarjuna and R. P. Rao, "Evaluation of forecasting methods from selected stock market returns," *Financial Innovation*, vol. 5, no. 1, pp. 1–16, 2019.
[9] X.-d. Zhang, A. Li, and R. Pan, "Stock trend prediction based on a new status box method and adaboost probabilistic support vector machine," *Applied Soft Computing*, vol. 49, pp. 385–398, 2016.