The application of artificial intelligence investment in capital markets: A case study of two constituent stocks of Dow Jones

Wei-Yuan Lin and Yuan-De Chu*

Department of Economics, Soochow University, Taipei, Taiwan.

Article History
Received 02 March, 2020
Received in revised form 13 April, 2020
Accepted 15 April, 2020

Keywords:
Reinforcement learning,
Technical indices,
Random forest,
SVM,
Prediction accuracy,
Online investment strategy.

Article Type:
Full Length Research Article

ABSTRACT
With the continuous advancement of science and technology, people continue to think about how to develop from weak artificial intelligence (AI) to strong AI. Researchers today already have big data richer than in the past, and use many new ideas of AI or data mining techniques to create many amazing achievements in various fields, such as the applications of sound, image and natural language. In addition, investment is also a hot topic that many people continue to explore, because investors always want to benefit from their decision-makings easily. In this paper, a reinforcement learning method is proposed to deal with the stocks market prediction. The information of the US stocks and data mining techniques were used to construct an online investment strategy, and conducted a comparative analysis of its different parameters. In this study, R software was apply to write the automated programs and build multiple technical indices or features. The shallow learning method--random forest (RF) was used to screen important features, and the support vector machine (SVM) classification method was also used to construct an online investment strategy. The empirical results obtained revealed that the average prediction accuracy can be increased significantly by adjusting the parameters of this model. Therefore, biology-based algorithms (GA, PSO and FOA etc.) or deep learning techniques (CNN, RNN etc.) can be further adopted in the future in order to obtain better accuracy and faster research results.

INTRODUCTION

With the development and progress of science and technology, people greatly improve computer computing performance. Coupled with the diversity and richness of information (data), that makes research and analysis become more attractive. Besides, due to the rapid development of artificial intelligence in recent years, it is now spread across all the domains of machine learning, such as the applications of sound, image and natural language or investment, etc. Now people can escape the limits of traditional statistical analysis, combining financial features to make investment strategies with computer algorithms, and get more diversified prediction methods and more accuracy prediction results.

In the securities exchange market, computers used to help human to trade have long existed. For example, Goldman Sachs uses their supercomputers for high-frequency trading. These trading strategies could make profit in a very short period of time. Although people have different opinions about using computers for transactions, but it is undeniable that the AI or data-mining technique shows its great potential in financial markets. Therefore, the main purpose of this paper is not simply to build an automated trading system, but also study how to analyze...
time series data more effectively.

The main structure of this article is as follows: The first section introduces the research motivation and objective, section 2 demonstrates research methods, section 3 applies the Reinforcement Learning and data mining techniques to construct an online investment strategy, and also conduct a comparative analysis of its different parameters, and section 4 will show our empirical research in US stocks. The paper ends with the research conclusion.

RESEARCH METHODS

In this section, we are going to introduce some of the concepts and mathematics behind Decision Trees (Kass, 1980) and RF (Breiman, 2018). Then SVM is also discussed.

Decision tree

Decision tree is a data mining technology commonly used in the field of machine learning (Crammer and Singer, 2002) technology in artificial intelligence. It can be regarded as an extended model of regression analysis (Wei-Yuan et al., 2018) and used for classification prediction. For example, based on the customer’s credit information, interest payment records and other information to predict whether to default and risk early warning (Wei-Yuan et al., 2018). The decision tree is to recursive partitioning a set of data according to different conditions at each stage. The biggest difference from regression analysis is that explanatory variables can be reused in different cutting stages.

Random forest (RF)

RF is mainly used in regression and classification (Liaw and Wiener, 2002). The prediction accuracy is improved without a significant increase in the amount of computation. It is a combination learning algorithm based on decision tree classifier and proposed by Breiman (2001) and Ho (1995). The classic RF classifier is a Classification and Regression Tree (CART), which combines learning through the Bagging algorithm, and randomly selects variables to perform attributes splitting when the CART tree grows. The main difference between RF and a decision tree is that when the RF performs a spanning tree, it has a good calculation method to prevent over-fitting.

RF is extracted from the data set and put back through repeated sampling. Different bootstrap training samples can be obtained, and then each classifier is trained. The number of samples in the sample set is assumed to be n, the number of Bootstrap samples drawn every time the extraction and return method is also n, then the probability that each sample is not drawn is \((1 - \frac{1}{n})^n\). It means that when n is large, the probability value is about 0.368, that is, about 37% of the samples in the original sample set are not selected when each sample is taken, and this unselected sample is called Out-of-Bag (OOB). It can be used as an internal estimate to calculate OOB values, and the training results of RF can also be used to evaluate the strength and correlation of forests.

Support vector machine (SVM)

SVM is a supervised machine learning method for data classification and pattern recognition. It was established in 1995 by Dr. Vapnik (Cortes and Vapnik, 1995), based on Structural Risk Minimization (SRM) in statistical learning theory. It promotes the correspondence between “independent variables” and “dependent variable” from the original lower-dimensional vector space to the higher-dimensional vector space.

In real life, almost every case we face is a problem of non-linear data classification. When the data is linearly
separable, the hyperplane is used for classification. But if you need to classify non-linear data, you can use Kernel Functions to change the data type. The main concept is to map the input data from the original Low Dimensional through the transformation of the Kernel Function to the High Dimensional (Φ: R^2 → F). In other words, the Kernel Function is to classify non-linear data into linear data as shown in the Figure 1. In the original input space, the data is non-linear separate, but through the process of mapping function, the data is projected into high dimensions to make the data linear, and make the data more dispersed with each other to facilitate grouping.

For example, suppose the mapping function from 2D projection to 3D is \(\phi: R^2 \rightarrow R^3\)

\[\phi(X) = \phi(x_1, x_2) = (z_1, z_2, z_3) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)\]  

Therefore, originally,

\[(0, 0) \rightarrow (0, 0, 0),\]
\[(1, 0) \rightarrow (0, 0, 1),\]
\[(0, 1) \rightarrow (0, 1, 0),\]
\[(1, 1) \rightarrow (1, \sqrt{2}, 1).\]

So when the data is more dispersed in \(R^3\), it will be easier to find the best hyperplane separated data as two parts.

There are four common Kernel Functions as follows:

- Linear: \(\langle \xi, x \rangle\)
- Polynomial: \(y \langle \xi, x \rangle + r\)
- RBF: \(\exp(-\gamma \| \xi - x \|^2)\)
- Sigmoid: \(\tanh(\gamma \langle \xi, x \rangle + r)\) where \(\gamma > 0\)

This method is based on the Wahba's theorem; any normal vector can be represented by a linear combination of other vector branches projected to a high dimension below:

\[f(\phi(\vec{X})) = \sum_{i=1}^{N} \alpha_i \phi(\vec{X}_i)\phi(\vec{X}) + b\]

\[= \sum_{i=1}^{N} \alpha_i K(\vec{X}_i, \vec{X}) + b\]  

(2)

Where \(K(\vec{X}_i, \vec{X}) = \phi(\vec{X}_i)\phi(\vec{X})\). Therefore, Equation 2 becomes a linear combination of Kernel Function.

**REINFORCEMENT LEARNING AND ONLINE INVESTMENT STRATEGY**

In this paper we use "R software" to write Online Investment Strategy programs based on the concept of Reinforcement Learning. During the overall research and analysis, some papers of the "RF" (Breiman, 2001; Han and Mao, 2014; Ho, 1995; Liaw and Wiener, 2002; Segal, 2004) are for references. We use the RF method to select several important technical indices for the strategies. The main strategies in our stock exchange are the following three, "Buy, Hold and Sell". In reference to relevant "Decision Classification" technology and theoretical findings, "SVM" (Bennett and Campbell, 2000; Cortes and Vapnik, 1995; Han and Mao, 2014; Karazmodeh et al., 2013; Platt, 1999; Rong-En et al., 2005; Scholkopf et al., 2000; Wei-Yuan et al., 2018) are also cited. Another related researches are also the application of “Machine Learning” in the fields of Investment and Prediction (Han and Mao, 2014; Karazmodeh et al., 2013; Shen et al., 2012) in this paper.

In this paper, the Reinforcement Learning Method and Online Investment Strategy is shown in Figure 2. First, we will get the stock daily data from Web, that is, Open, Close, High, Low and Volume and the adjusted price (Appendix I). Then, take the average value of "Close price, Low price, and High price" as the Average Price \(\overline{P}_i\) in Equation 3.

\[\overline{P}_i = \frac{P_i + L_i + H_i}{3}\]  

(3)

At the same time, we calculate ten technical indices (features) shown in Table 1. Next we also calculate the rate of return, which represents the percentage change of the average price below:

\[V_i = \{\frac{P_{i+1}-P_i}{P_i}\}_{i=1,2,...,k}\]  

(4)

Where \(\overline{P}_i\) represents the average price on day \(i\). We believe that the transaction will only be conducted if the benefits obtained during the transaction need to exceed the transaction costs (p%), otherwise the transaction will be abandoned. Next, we create a new indices (Appendix II), the cumulative k-day accumulated return is calculated as follow:

\[T_i = \sum_{j=1}^{k} V_j \text{ such that } |V_j| > p\%\]  

(5)

Then we divide into three parts to construct the following Trading strategy (Table 2).

Finally, we use RF method to select the more important features. And then think of technical indices as "explanatory variables" and \(T\) as "explained variable", SVM method is used to find the optimal strategy. Effective strategy parameters \((a1, a2)\) selections are also considered in this paper.
**Figure 2.** Flow chart of reinforcement learning and online investment strategy.

**Table 1.** Main technical indices.

<table>
<thead>
<tr>
<th>Index</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATR</td>
<td>True Range/Average True Range</td>
</tr>
<tr>
<td>ADX</td>
<td>Average Directional Index</td>
</tr>
<tr>
<td>MACD</td>
<td>Moving Average Convergence and Divergence</td>
</tr>
<tr>
<td>SAR</td>
<td>Parabolic Stop-and-Reverse</td>
</tr>
<tr>
<td>SMI</td>
<td>Stochastic Momentum Index</td>
</tr>
<tr>
<td>BBands</td>
<td>Bollinger Band</td>
</tr>
<tr>
<td>CHV</td>
<td>Chaikin Volatility</td>
</tr>
<tr>
<td>CLV</td>
<td>Close Location Value</td>
</tr>
<tr>
<td>MFI</td>
<td>Mooney Flow Index</td>
</tr>
<tr>
<td>Aroon</td>
<td>Aroon Index</td>
</tr>
</tbody>
</table>

These technical indices can be conveniently obtained in TTR of R package.

**Table 2.** Trading strategy.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>S</td>
</tr>
<tr>
<td>0</td>
<td>H</td>
</tr>
<tr>
<td>+1</td>
<td>B</td>
</tr>
</tbody>
</table>

Note: -1 means Sell, 0 means Hold, and 1 means Buy.

**EMPIRICAL STUDY**

This research collected the past two decades (2000~2019) stock data, two constituent stocks of Dow Jones Industrial Average (DJIA)-- and CAT. It covers dotcom bubble in 2001, the global financial crisis in 2008 and the Sino-US trade war in 2019. It has also experienced many stock market fluctuations and boom cycles.

In this paper, we use "R software" to write programs, and automatically obtain daily stock price information directly from the network. At the same time, the data mining method is used for investment forecasting. Get higher rewards with automated transaction analysis.
systems and decide when to buy or sell the stock during the research period of time. We also analyze two different accuracy criteria ("SHB" and "SB") by two methods ("Fixed Value Method, and Interval Parameter Method").

In this empirical study, based on "SHB" strategy, the best accuracy (Accuracy 1) using the "Fixed Value Method" falls between "45~58%". On the other hand, based on "Interval Parameter Method" (Accuracy 2) can be raised to between "62~68%". In other words, we conclude that the "Interval Parameter Method" is better than the "SHB" strategy; Similarly, based on "SB" strategy, the best accuracy (Accuracy 1) using the "Fixed Value Method"
Figure 5. Technical indices of AAPL.

Figure 6. Technical indices of CAT.

Figure 7. The IncNode Purity of important features.

importance(rf@fitted.model)
                      IncNodePurity
myTTR.stock.myATR         141.05243
myTTR.stock.mySMI          94.61538
myTTR.stock.myADX         109.45003
myTTR.stock.myAroon       68.46017
myTTR.stock.myBBands      84.80648
myTTR.stock.myChaikin     52.93581
myTTR.stock.myCLV         62.82205
myTTR.stock.myMACD        125.78313
myTTR.stock.myMFI         83.00895
myTTR.stock.mySAR         164.69148
Figure 8. Accuracy 1 and Accuracy 2 for AAPL.

Figure 9. Accuracy 1 and Accuracy 2 for CAT.

Table 3. Confusion matrix.

<table>
<thead>
<tr>
<th>Investment strategy</th>
<th>Signal. True:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Buy</td>
<td>Hold</td>
<td>Sell</td>
<td></td>
</tr>
<tr>
<td>Signal. predict:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Buy</td>
<td>BB</td>
<td>BH</td>
<td>BS</td>
<td></td>
</tr>
<tr>
<td>Hold</td>
<td>BH</td>
<td>HH</td>
<td>HS</td>
<td></td>
</tr>
<tr>
<td>Sell</td>
<td>SB</td>
<td>SH</td>
<td>SS</td>
<td></td>
</tr>
</tbody>
</table>
Table 4. The overall information period is from 2000/1/1 to 2019/10/25.

<table>
<thead>
<tr>
<th>Stock</th>
<th>US stocks (DJIA)</th>
<th>AAPL (Apple)</th>
<th>CAT (Caterpillar)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training period</td>
<td>2000-01-01–2018-12-31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research method</td>
<td>Technical indicators, SVM, decision tree (RF method)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data source</td>
<td>Financial stock market website</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Selected important technical indices.

<table>
<thead>
<tr>
<th>Stock</th>
<th>US stocks (DJIA)</th>
<th>AAPL (Apple)</th>
<th>CAT (Caterpillar)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Important(<a href="mailto:randomForest_Tag@fitted.model">randomForest_Tag@fitted.model</a>)&gt;80</td>
<td>myTTR.stock.mySAR</td>
<td>myTTR.stock.mySAR</td>
<td>myTTR.stock.mySAR</td>
</tr>
<tr>
<td>Very important</td>
<td>myTTR.stock.myATR</td>
<td>myTTR.stock.myATR</td>
<td>myTTR.stock.myATR</td>
</tr>
<tr>
<td>technical indicators</td>
<td>myTTR.stock.myMACD</td>
<td>myTTR.stock.myMACD</td>
<td>myTTR.stock.myMACD</td>
</tr>
<tr>
<td></td>
<td>myTTR.stock.myADX</td>
<td>myTTR.stock.myADX</td>
<td>myTTR.stock.myADX</td>
</tr>
<tr>
<td></td>
<td>myTTR.stock.mySMI</td>
<td>myTTR.stock.mySMI</td>
<td>myTTR.stock.mySMI</td>
</tr>
<tr>
<td></td>
<td>myTTR.stock.myBBands</td>
<td>myTTR.stock.myBBands</td>
<td>myTTR.stock.myBBands</td>
</tr>
<tr>
<td></td>
<td>myTTR.stock.myMFI</td>
<td>myTTR.stock.myMFI</td>
<td>myTTR.stock.myMFI</td>
</tr>
</tbody>
</table>

Table 6. Accuracy 1 and Accuracy 2 for AAPL.

<table>
<thead>
<tr>
<th>N:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Avg.</th>
<th>Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy1:SBH</td>
<td>0.4133</td>
<td>0.4388</td>
<td>0.4184</td>
<td>0.4541</td>
<td>0.3316</td>
<td>0.2449</td>
<td>0.2755</td>
<td>0.2653</td>
<td>0.3214</td>
<td>0.3622</td>
<td>0.35</td>
<td>0.0048</td>
</tr>
<tr>
<td>Accuracy2:SBH</td>
<td>0.6786</td>
<td>0.5612</td>
<td>0.6327</td>
<td>0.5000</td>
<td>0.5102</td>
<td>0.4949</td>
<td>0.5561</td>
<td>0.4490</td>
<td>0.4439</td>
<td>0.4694</td>
<td>0.53</td>
<td>0.0049</td>
</tr>
<tr>
<td>Accuracy1:SB</td>
<td>0.4260</td>
<td>0.4624</td>
<td>0.4239</td>
<td>0.4579</td>
<td>0.3316</td>
<td>0.2487</td>
<td>0.2789</td>
<td>0.2604</td>
<td>0.3249</td>
<td>0.3590</td>
<td>0.36</td>
<td>0.0054</td>
</tr>
<tr>
<td>Accuracy2:SB</td>
<td>0.5625</td>
<td>0.5057</td>
<td>0.5267</td>
<td>0.5188</td>
<td>0.4240</td>
<td>0.3065</td>
<td>0.3717</td>
<td>0.3140</td>
<td>0.3817</td>
<td>0.4074</td>
<td>0.43</td>
<td>0.0069</td>
</tr>
</tbody>
</table>

Table 7. Accuracy 1 and Accuracy 2 for CAT.

<table>
<thead>
<tr>
<th>N:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Avg.</th>
<th>Var.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy1:SBH</td>
<td>0.3571</td>
<td>0.3469</td>
<td>0.4541</td>
<td>0.5408</td>
<td>0.4796</td>
<td>0.5561</td>
<td>0.5459</td>
<td>0.5714</td>
<td>0.5612</td>
<td>0.5765</td>
<td>0.50</td>
<td>0.0062</td>
</tr>
<tr>
<td>Accuracy2:SBH</td>
<td>0.6224</td>
<td>0.5306</td>
<td>0.4643</td>
<td>0.5867</td>
<td>0.5102</td>
<td>0.5765</td>
<td>0.5612</td>
<td>0.5867</td>
<td>0.5765</td>
<td>0.6020</td>
<td>0.56</td>
<td>0.0018</td>
</tr>
<tr>
<td>Accuracy1:SB</td>
<td>0.3557</td>
<td>0.3579</td>
<td>0.4611</td>
<td>0.5521</td>
<td>0.4821</td>
<td>0.5648</td>
<td>0.5487</td>
<td>0.5744</td>
<td>0.5691</td>
<td>0.5833</td>
<td>0.50</td>
<td>0.0063</td>
</tr>
<tr>
<td>Accuracy2:SB</td>
<td>-</td>
<td>0.3622</td>
<td>0.4667</td>
<td>0.5975</td>
<td>0.5057</td>
<td>0.5786</td>
<td>0.5590</td>
<td>0.5897</td>
<td>0.5820</td>
<td>0.5969</td>
<td>0.54</td>
<td>0.0051</td>
</tr>
</tbody>
</table>

Value Method” falls between "46~58%". And “Interval Parameter Method (Accuracy 2) can be improved to “56~60%”. It also shows that the “Interval Parameter Method” still has better result. For more detail, please refer to Tables 3 to 7 and Figures 3 to 9.

Parameters setting and other information explained

This empirical study includes the following six items in R program below:
(1) Stock: Two constituent stocks of Dow Jones:
(2) Corporation (AAPL) and American Caterpillar (CAT).
(3) Main "Strategies": "SHB and SB" two ways.
(4) "Accuracy" calculation method:

SHB: Accuracy function= (SS+HH+BB)/(S+H+B)

SB: Accuracy function= (SS+BB)/(S+B)

Where, SS, Real stocks are down and the forecast signal is sell; HH, real stock is held and the forecast signal is held; BB, real stocks are up and the forecast signal is buy; S, the forecast signal is a sell regardless of whether the real stock is up or down or held; H, the forecast signal is held regardless of whether the real stock is up or down; B, Regardless of whether the real stock is up or down or held, the forecast signal is buy. (Note: The first "S" in "SS" is the predicted signal, and the second "S" is the real signal.)

The "SHB" (or 2) method is used to calculate the accuracy. For two stocks, different a1 and a2 parameters are used at the same time, and N (0 ~ 10) technical variable selections are used for comparison.

These two kinds of "Fixed Value Method" and "Interval Parameter Method" are shown as following:

(A) Fixed value method: The a1 and a2 are given a fixed value to calculate the accuracy (a1 = 0.0095, a2 = 0.0095).

(B) Interval parameter method:
(a) Parameters a1 (where: a1 = -a2) gives a range (0.001 ~ 0.3).
(b) Parameters a1 and a2 perform accuracy calculations in this interval to produce the most relevant combination.

(5) "N": The number of variables expected to be added in the technical indices function (TTR): The range is 1 ~ 10.

(6) p = 0.025: p represents transaction cost; for the basic requirement of stock trading, the value is currently set at 2.5%.

(7) k = 10:k is the period of time planned for each transaction. Here we set to 10.

Empirical results of AAPL and CAT

The training period is from January 1, 2000 to December 31, 2018. The test period is from January 1, 2019 to October 25, 2019 (Table 4).

Analysis results: Screening technical indices

Regarding AAPL and CAT, the important technical indices of RF. The final selection sort has more than 90% of the same. It shows that the selected relevant important indices have certain reference value. In terms of importance (node purity IncNodePurity), although the technical indices of the AAPL and CAT are the same; however, the importance is still different (take AAPL as an example, there are seven technical indices selected for Importance> 80, but only three are selected by CAT). And the important features selected by RF in R is also shown Figures 5 to 7.

For technical indices, this "IncNodePurity" value is produced according to its degree of importance based on RF criteria as shown in the Figure 7. The best important feature is "SAR" (164.69148), and followed by "ATR" (142.05243), etc. In this paper, I only show important features in my model and try to find the best result of accuracy.

Analysis results: Performance evaluation (Appendix III)

Case 1: AAPL

1. Strategy 1 (SHB) has better accuracy than Strategy 2 (SB).
2. For the number of technical indices (N), N = 4 is preferred.
3. In Strategy 1 (SHB), using Accuracy 2 (53%) and Accuracy 1 (35%) after empirical testing, the average prediction accuracy can be improved by about 18%. This result shows that the average prediction accuracy can be increased significantly by adjusting parameters of our model (Table 6 and Figure 8).

Case 2: CAT

1. Strategy 1 (SHB) has better accuracy than Strategy 2 (SB).
2. For the number of technical indices (N), N = 4 is preferred.
3. In Strategy 1 (SHB), using Accuracy 2 (56%) and Accuracy 1 (50%), the average prediction accuracy can be improved 6%. This result also shows that the average prediction accuracy can be increased significantly by adjusting parameters of our model (Table 7 and Figure 9).

Empirical results: Heat map and Important technical indices selected

For the Interval Parameter Method, the parameter a2 (where: a1 = -a2) is set in a range (0.001 ~ 0.3). Based on four important features (SAR, ADX, MACD, and ATR) selected by RF, Heat Maps of AAPL and CAT by SVM are shown in Figures 10 to 13 respectively.
Figure 10. A, "Accuracy" heat map; B, technical indices (N = 4).

Figure 11. A, "Accuracy" heat map; B, technical indices (N = 4).
Figure 12. A, "Accuracy" heat map; B, technical indices (N = 4).

Figure 13. A, "Accuracy" heat map; B, technical indices (N = 4).
CONCLUSION

This paper uses "two stocks (AAPL and CAT)", "two strategies (SHB and SB)", "two precision calculation method (Fixed Value Method and Interval Parameter Method)", "N (= 1 ~ 10) "technical index", return rate parameter (p = 0.025) and "accumulated observation days (k =10 days)” to analyze. Therefore, there are 80 combinations of cases that can be analyzed and compared.

- After empirical backtesting, if the accuracy is calculated using the “SHB” strategy, the best accuracy (Accuracy1) for using the “Fixed Value Method” falls between "45~58%". If the Interval Parameter Method is used instead, the accuracy (Accuracy 2) will fall between "62~68%", which also shows that "Interval Parameter Method" with better accuracy.
- If the accuracy is calculated using the “SB” strategy, the best accuracy (Accuracy 1) using the “Fixed Value Method” falls between "46~58%". If the “Interval Parameter Method” is used, the accuracy (Accuracy 2) falls between "56~60", which also shows that the “Interval Parameter Method” also has better accuracy.
- In addition, the accuracy calculation methods of two different strategies, "SHB" and "SB", are compared, and the "Interval Parameter Method" is also used to show that the "SHB" strategy will have better accuracy.
- It can be seen from the K-line chart during the training period, AAPL and CAT stocks have shared the same increasing trend. Compared with the accuracy of the test period, AAPL is higher than CAT (68%> 62%) from our empirical results. The possible reason could be that AAPL’s stock price has an upward trend during this test period, but CAT occurs downward trend.

Findings of our study

- According to Han and Mao (2014) the accuracy formula used is "SB", which is mainly different from "SHB" ignoring “Hold” on the overall accuracy calculation. Therefore, the first analysis of our study is to add the "SHB" method to compare the accuracy of the two differences. The results show that the "SHB" method can indeed obtain “better” accuracy.
- In Han and Mao (2014) program, “Testing data” and “Training data” overlap in model training. The mistake in writing this investment program is that the “data deadline” is not correctly set to “asdate”, and eventually occurs the accuracy error (“overestimate”).
- Comparing the “a1 and a2” parameters using the difference between the “Fixed Value Method” and the “Interval Parameter Method”. The use of the more flexible “Interval Parameter Method” setting can indeed obtain better accuracy.

- For the selection of the main technical indices of the two stocks, the first three main indices have selected “SAR, ATR and MACD”, showing that these indices have a better explanatory power for the analysis results. Our finding can be used as a reference for the analysis of stock market trends.
- This paper uses artificial intelligence methods (SVM, RF) for analysis and research. The analysis process and related results can be used as an immediate reference for investors, and to forecast stock trends and evaluate the optimal time for stock trading.

Research contributions and features

- Because in practice, the stock transaction should include "Sell/Hold/Buy" three possible transactions. Therefore, objectively adding "Hold" decision-making behavior is more reasonable in the real world. One of the contributions of our research is confirmed after empirical evidence that the "SHB" method has better "accuracy".
- For the selection (screening) of the "best variables" and for loading the daily transaction data of individual stocks via "WEB" retrieval. This "automated" processing enhances the "instant decision-making and analysis" required for data analysis and trading transactions, which is another contribution of this research.
- At present, related research is mostly in fields of sound, image, natural language or medical study, and only a few people propose the online investment strategies.

Future research directions

- In addition to using the “Interval Parameter Method” for the a1 and a2 parameters, “biology-based algorithms [GA (Scholkopf et al., 2000), PSO (Evers and Ghalia, 2009; Konstantinos Parsopoulos and Michael, 2002), FOA (Wei-Yuan, 2016) etc.] or Deep Learning Techniques [CNN,RNN (Gulli and Pal, 2017) etc.] can be further adopted in order to obtain better accuracy and research results more faster in the future.
- For other model parameters (p, k, and n) can also be adjusted to make the research results more pragmatic and accurate.
- Although this paper belongs to shallow learning-- SVM and RF, this research on “less data” is still operational quit good. For more research and analysis of “big data, we canal so consider the “Deep Learning” and “Reinforcement Learning” techniques in our model.

REFERENCES

Appendix

Appendix I. The following shows the fundamental statistics of stock prices and volume.

a) Stock: AAPL

<table>
<thead>
<tr>
<th>Index</th>
<th>AAPL.Open</th>
<th>AAPL.High</th>
<th>AAPL.Low</th>
<th>AAPL.Close</th>
<th>AAPL.Volume</th>
<th>AAPL.Ajusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.  :2000-01-03 Min.  :0.9279</td>
<td>Min.  :0.9421</td>
<td>Min.  :0.9086</td>
<td>Min.  :0.9371</td>
<td>Min.  :9.835e+06</td>
<td>Min.  :0.8135</td>
<td></td>
</tr>
<tr>
<td>1st Qu.:2004-12-17 1st Qu.:4.7093</td>
<td>1st Qu.:4.8036</td>
<td>1st Qu.:4.6186</td>
<td>1st Qu.:4.6757</td>
<td>1st Qu.:4.0587</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean :2009-11-29 Mean :57.8580</td>
<td>Mean :58.1298</td>
<td>Mean :57.0194</td>
<td>Mean :57.5887</td>
<td>Mean :1.158e+08</td>
<td>Mean :53.7237</td>
<td></td>
</tr>
<tr>
<td>3rd Qu.:2014-11-11 3rd Qu.:97.7900</td>
<td>3rd Qu.:98.7100</td>
<td>3rd Qu.:97.5686</td>
<td>3rd Qu.:91.524e+08</td>
<td>3rd Qu.:90.4739</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b) Stock: CA

<table>
<thead>
<tr>
<th>Index</th>
<th>CAT.Open</th>
<th>CAT.High</th>
<th>CAT.Low</th>
<th>CAT.Close</th>
<th>CAT.Volume</th>
<th>CAT.Ajusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.  :2000-01-03 Min.  :15.03</td>
<td>Min.  :15.06</td>
<td>Min.  :14.50</td>
<td>Min.  :14.91</td>
<td>Min.  :776400</td>
<td>Min.  :7.866</td>
<td></td>
</tr>
<tr>
<td>1st Qu.:2004-12-17 1st Qu.:38.70</td>
<td>1st Qu.:39.27</td>
<td>1st Qu.:38.10</td>
<td>1st Qu.:38.63</td>
<td>1st Qu.:3760800</td>
<td>1st Qu.:25.312</td>
<td></td>
</tr>
<tr>
<td>Mean :2009-11-29 Mean :70.96</td>
<td>Mean :71.72</td>
<td>Mean :70.95</td>
<td>Mean :606426</td>
<td>Mean :56.818</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Qu.:2014-11-11 3rd Qu.:93.06</td>
<td>3rd Qu.:93.94</td>
<td>3rd Qu.:92.92</td>
<td>3rd Qu.:92.92</td>
<td>3rd Qu.:7163500</td>
<td>3rd Qu.:75.487</td>
<td></td>
</tr>
</tbody>
</table>

Appendix II. The following shows the accumulated return (T) for the stocks AAPL and CAT.

a) Stock: AAPL

<table>
<thead>
<tr>
<th>Index</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.  :2018-01-02 Min.  :-0.94988</td>
<td>1st Qu.:2018-04-03 1st Qu.: -0.21787</td>
</tr>
<tr>
<td>Median:2018-07-02 Median : 0.00000</td>
<td>Mean :2018-07-01 Mean : -0.01824</td>
</tr>
<tr>
<td>3rd Qu.:2018-09-29 3rd Qu.: 0.13394</td>
<td>Max. :2018-12-31 Max. : 1.08543</td>
</tr>
<tr>
<td>NA's : 10</td>
<td></td>
</tr>
</tbody>
</table>

b) Stock: CAT

<table>
<thead>
<tr>
<th>Index</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.  :2018-01-02 Min.  :-1.43217</td>
<td>1st Qu.:2018-04-03 1st Qu.: -0.23315</td>
</tr>
<tr>
<td>Median:2018-07-02 Median : 0.00000</td>
<td>Mean :2018-07-01 Mean : -0.05344</td>
</tr>
<tr>
<td>3rd Qu.:2018-09-29 3rd Qu.: 0.14596</td>
<td>Max. :2018-12-31 Max. : 0.97175</td>
</tr>
<tr>
<td>NA's : 10</td>
<td></td>
</tr>
</tbody>
</table>
Appendix III. The following shows the strategy parameters (a1 and a2) and the confusion matrix under two methods.

a) Stock: AAPL

(1) "Fixed value method"

```
> cbind(a1[ind[,1]],a2[ind[,2]])
[,1] [,2]
[1,] -0.2877959 0.3
[2,] -0.2938980 0.3
[3,] -0.3000000 0.3
```

```
> table(signal.p,signal.true)
signal.true
signal.p Buy Hold Sell
Buy 68 18 24
Hold 7 4 1
Sell 56 8 10
```

```
> accuracy(signal.p,signal.true)
[1] 0.4183673
```

(2) "Interval parameter method"

```
> cbind(a1[ind[,1]],a2[ind[,2]])
[,1] [,2]
[1,] -0.001 0.01930612
[2,] -0.001 0.01930612
```

```
> table(signal.p,signal.true)
signal.true
signal.p Buy Hold Sell
Buy 48 10 24
Hold 3 1 0
Sell 33 13 64
```

```
> accuracy(signal.p,signal.true)
[1] 0.5765306
```

b) Stock: CAT

(1) "Fixed value method"

```
> cbind(a1[ind[,1]],a2[ind[,2]])
[,1] [,2]
[1,] -0.001 0.01930612
```

```
> table(signal.p,signal.true)
signal.true
signal.p Buy Hold Sell
Buy 48 7 23
Hold 0 4 1
Sell 35 12 66
```

```
> accuracy(signal.p,signal.true)
[1] 0.602408
```

(2) "Interval parameter method"