Stock Market Investment Using Machine Learning

Chen Chen

Centre for Computational Finance and Economic Agents
University of Essex

This dissertation is submitted for the degree of
Doctor of Philosophy

University of Essex December 2022

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Chen Chen
December 2022

Acknowledgements

I would like to express sincere gratitude to supervisor Dr. Edward P. K. Tsang. I have been his student since I was a master student in CCFEA. His academic foresight always inspires me. I would also appreciate his care when I encountered a difficult time due to personal reasons. I sincerely hope that he can have a happy retirement life.

I want to thank supervisor Dr. Carmine Ventre, who guided me on the specific details of machine learning models.

I must express my great respect and gratitude to supervisor Dr. Michael Fairbank. He spent much time reading my thesis and gave me professional and meticulous guidance, both in terms of experimental results and presentation of the thesis. Through weekly online meetings and email exchanges, we have discussed and revised each thesis chapter in depth. I have benefited greatly from his professionalism, diligence, and patience, and his work style of attention to detail and efficiency is worth learning from.

I need to appreciate all of the teachers who taught me in 2012-2013 during my study in CCFEA: Dr. John O'Hara. He is my teacher of the course CF961 and CF966 and the chairman of my Ph.D. program. All of the CCFEA students would admire his dedication to his career. Dr. Wing Lon Ng, professionally trained us in Matlab skills, and he is strict but very responsible. I cannot mention all of my teachers' names, but I am grateful.

I would like to appreciate two friends:

Li Shengnan is my colleague, and he sheltered me when I could not find proper accommodation. Besides, he offered me his 44 cores 88 threads computer to acquire my data quickly. Wang Ziyao, graduated from the University of Essex in 2017. They are my best friends in my life.

At last, I need to thank my parents, their support in my postgraduate makes me enjoy the happiness of studying for many years. In particular, I want to thank my father, who independently assumed financial support during my Ph.D.

Regardless of the outcome, the experience during my Ph.D., especially the period between the first and the second VIVA, will be the most precious asset in my life.

Abstract

Genetic Algorithm-Support Vector Regression (GA-SVR) and Random Forest Regression (RFR) were constructed to forecast stock returns in this research. 15 financial indicators were selected through fuzzy clustering from 42 financial indicators, then combined with 8 technical indicators as input space, the 10-day stocks return was used as labels. The results show that GA-SVR and RFR can make compelling forecasting and pass the robustness test. GA-SVR and RFR exhibit different processing preferences for features with different importance. Furthermore, by testing stock markets in China, Hong Kong (China) and the United States, the model shows different effectiveness.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background

The development of machine learning provides suitable technical tools for predicting stock prices. Predicting stock prices usually requires reference to multi-dimensional time-series data, such as stock price, technical indicators, financial indicators, national economic indicators. The panel data matrix composed of these data is called feature space in machine learning. When the feature space's dimensionality (the number of features) is high, the traditional statistical inference methodology is generally ineffective, while machine learning becomes an appropriate choice. This research will apply random forests and support vector machines to predict stock price.

Are stock prices predictable? It might be the most attractive question in the financial area. In academia, there is a general awareness that the question is closely related to the stock market efficiency.

The research on market efficiency stems from people's predictive research on capital market price. Some scholars found that stock prices were unpredictable and show random walk [18]. Based on reviewing the random walk theories and empirical research, Fama in 1970 supported the randomness of the stock price with compelling evidence and defined 'efficient market' for the first time: a market that can quickly adjust to new information [19]. Fama in 1991 changed the efficient market definition to a market where assets fully reflect all available information, which means that the market can be considered efficient if all helpful information is reflected in the prices of securities in an unprejudiced way. Fama divided the efficient market hypothesis into three levels: weak-form efficiency, semi-strong form efficiency and strong-form efficiency [20].

Most scholars' early empirical research supported the efficient market hypothesis, believing that the mature securities market is weakly efficient. However, as many market anomalies appeared, people began to doubt the effectiveness of the market.

The active investing industry should not exist if the efficient market hypothesis holds, but the active investment industry, in reality, has existed for a long time. Hence, this research will provide new evidence for the debate on market effectiveness in addition to predicting stock prices through machine learning.

1.2 Research Purpose

This thesis divides the research purpose into three sub-purposes.

1. Answer whether the stock price can be predicted by machine learning and explain the anomalies found in this process.

2. Compare the advantages and disadvantages of Random Forests and Support Vector Machines in applying stock price prediction.

3. Compare the market efficiency of the stock markets in different countries and regions.

The primary research purpose of this paper is to verify whether stock returns can be predicted by machine learning, which is research purpose 1. Since this thesis adopts two forecasting models: SVR and RFR, we can compare the advantages and disadvantages of the two models, which is purpose 2. The academic concept of prediction of individual stock returns comes from the efficient market hypothesis, that is, investors can obtain excess returns through active prediction. Hence the result of running prediction on different markets could help us to achieve purpose 3. In Chapter 8 of the thesis, we intend to apply the forecasting algorithm to three stock markets, if it generates significantly different results, then we can cautiously conclude: under this test, one market's efficiency is different from others'.

1.3 Overview of Core Model of the Thesis

This thesis is an interdisciplinary outcome of finance and machine learning. The methodology involves multiple machine learning fields such as fuzzy clustering, genetic algorithm parameter optimisation, random forests and support vector machines. Directly reviewing the technical details of the thesis will probably cause confusion. This section therefore provides an overview of the core model.

The core model of the thesis is a supervised machine learning model (either SVMs or RFs). We will use five figures to demonstrate the core methodology of the thesis: Figure 1.1 is the simplified process of supervised machine learning. Figure 1.2 is the demonstration of rolling window method. Figure 1.3 is the extension of Figure 1.1 with more details, Figure 1.4 and 1.5 are the extension of Figure 1.3 with specific models (SVR and RFR) of the thesis. In Figure 1.4 and 1.5, the author has marked the correspondence between the methodology structure and the chapters of thesis.

Machine learning can be divided into supervised learning and unsupervised learning. Supervised learning trains existing training samples to obtain an optimal model and then use this model to map inputs to generate corresponding outputs. The training samples of supervised learning contain both features and labels, while the labeling information of the training samples of unsupervised learning is unknown. Unsupervised learning aims to reveal the intrinsic properties and laws of the data through the learning of unlabeled training samples. In our methodology, Fuzzy Clustering (for feature engineering) is a typical unsupervised learning method. Both RFs and SVMs belong to supervised machine learning. Supervised machine learning can be simplified to the steps in Figure 1.1:

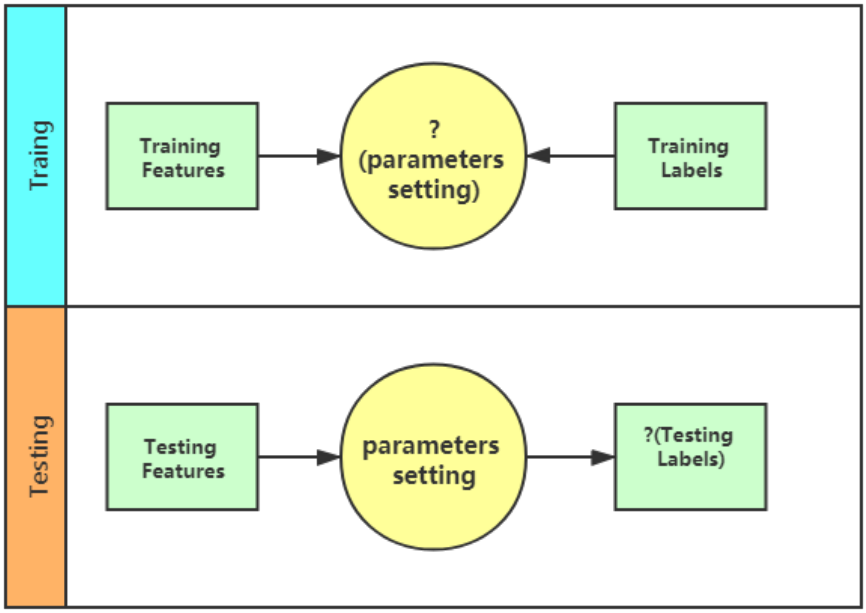


Fig. 1.1 Simplified process of supervised machine learning

1. In the training stage, input features and labeled results are known, we wish to to determine appropriate parameters.

2. In the testing stage, input features and parameters are known, we wish to determine labels.

In this thesis, when we apply the above method to time series prediction, the 'rolling window method' is introduced in Figure 1.2:

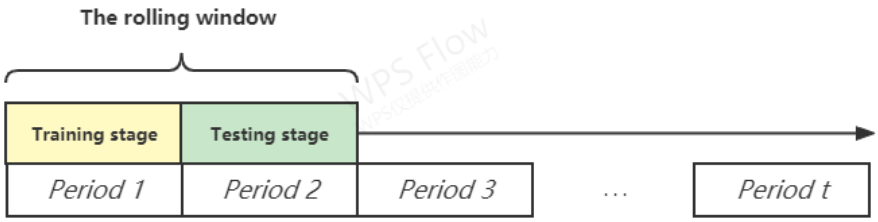


Fig. 1.2 The rolling window method

The mechanism of the rolling window method is:

- Step 1 Cut the time series data to equal periods.
- Step 2 Take period 1 as the training stage and period 2 as the testing stage.
- Step 3 Take period 2 as the training stage and period 3 as the testing stage.
- Repeat the rolling process of step 2 and step 3 until the last period then the rolling window method is completed.

Through the rolling window method, the SVMs and RFs are applied to the entire time series data.

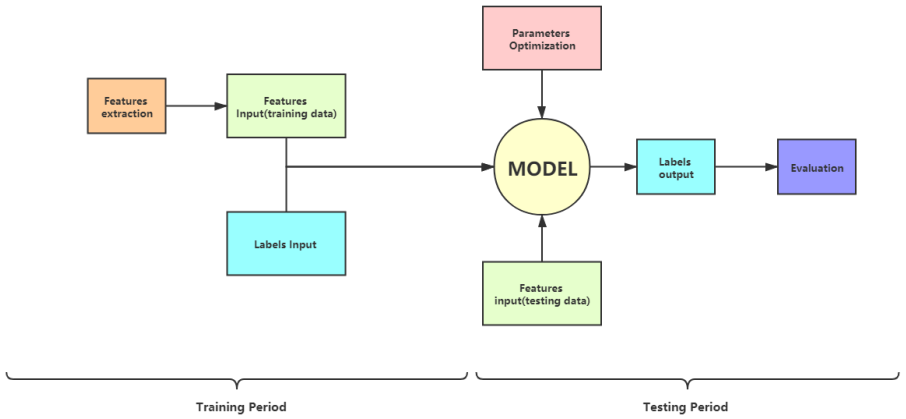


Fig. 1.3 Process of supervised machine learning

Figure 1.3 is the detailed process of the supervised machine learning algorithm. In the training period, to reduce the dimensionality and increase the validity of the input data, the input features could be extracted and constructed from the original data. It is called Feature Engineering. Different machine learning algorithms could apply various optimisation methods to find the appropriate parameter set (in this thesis, the SVR algorithm uses genetic algorithms for optimisation), which is called parameter optimisation.

After labeling the testing data (or obtaining value, depending on whether it is a classification algorithm or a regression algorithm), we need to evaluate the algorithm's effectiveness with the actual label/value. This step is called evaluation.

1.4 Thesis Structure

The dissertation is organized according to the structure in Figure 1.4 and 1.5. The corresponding chapters of the dissertation are marked. Chapter 3 is feature engineering. Chapter 4 introduces the regression algorithm based on support vector machine. Chapter 5 introduces the genetic algorithm applied to optimise support vector machine parameters. Chapter 6 introduces the regression algorithm based on random forest. Chapter 7 are the results and evaluation, which include the results and analysis of the models, the robustness test and the trading simulation. Chapter 8 compares the model's effectiveness in the United States' and Chinese stock markets. Chapter 9 is the conclusion.

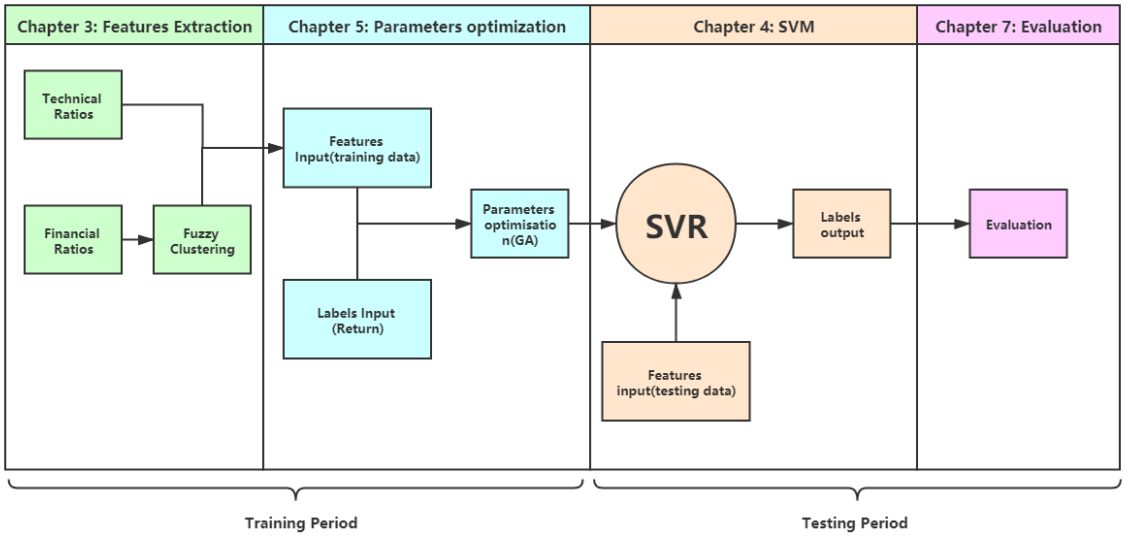


Fig. 1.4 Major chapters in the thesis and how they relate to the machine learning (SVR) pipeline

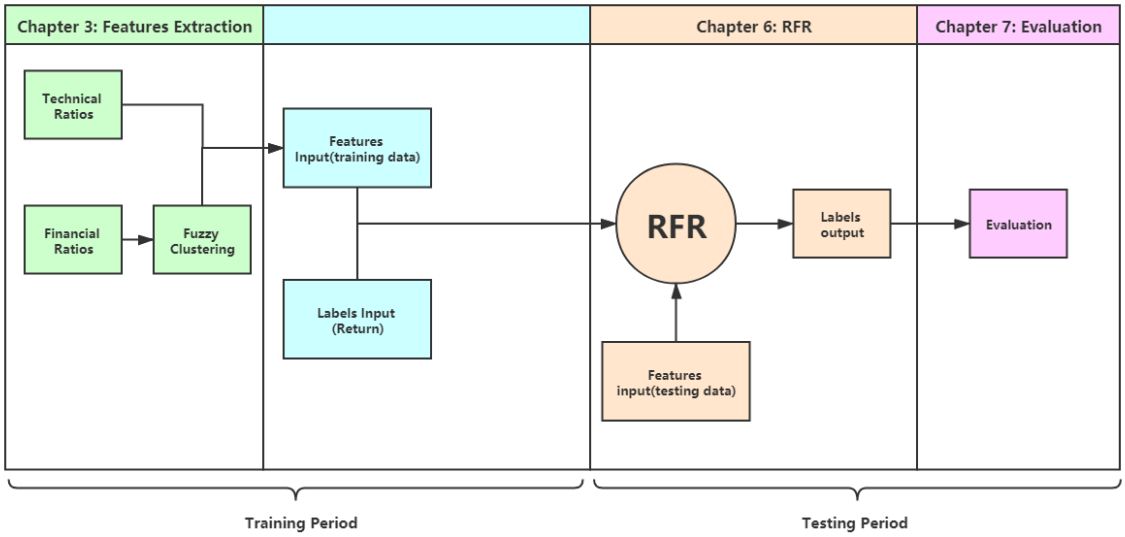


Fig. 1.5 Major chapters in the thesis and how they relate to the machine learning (RFR) pipeline

Chapter 2

Literature Review

The literature review is divided into three parts: section 2.1 Literature review on machine learning prediction, section 2.2 Literature review on linear model prediction,and section 2.3 Supporting literature used to support the technical aspects of the prediction model of the thesis including FCA features reduction and GA parameter optimisation.

The literature review will describe 15 papers in turn. Literature related to machine learning in section 2.1 will be reviewed and presented more detailed. For each paper that have a machine learning model, there will be a structured review followed by analysis which contrasts their work to the methodology used in this thesis. The structured review will highlight systematical core steps of the machine learning model, includes: a) input feature, labels and output. b) Feature engineering. c) Model. d) Evaluation. e) Result.

By analysing these five aspects, we can fully understand the pipeline of the methodology and how the papers are related to each other and to the work in this thesis. Therefore, each literature review about machine learning will analyse these aspects and then make discussions.

In addition, in section 2.5, we reproduced the experiment of paper 'Predicting the direction of stock market prices using random forest' [1]. Through the reproduction, we figured out the flaw of the original methodology, which will distort the prediction results. In this thesis we will avoid the flaw.

This literature review is not conventional, and has two main differences from conventional literature reviews:

1. Few literature reviews replicate and validate the methodology of others. However, the author believes that it is reasonable to maintain a moderate degree of skepticism about financial forecasting literature. We believe that the methods used by some financial forecasting papers to are unreasonable, and their results may be distorted [1–3, 7, 8].

One of the main mistakes is to equate time-series data with cross-sectional data, thereby assuming that security returns and volatility across time periods follow the same statistical distribution. Under this assumption, the mistake has three manifestations:

(a) Selecting the data forecast results for a specific time period and declaring that the forecast is valid.

(b) Using unused data in the training set as testing data instead of the real testing data.

(c) Having problems with over-fitting parameters

2. It can be concluded from the above discussion that readers may be misled by merely adopting the conclusions of the literature without delving into the underlying data and model construction. Therefore, in the literature review section, we adopt an unconventional structure, listing 15 papers and focusing on the details of their models' construction.

2.1 Literature Review on Machine Learning Prediction

There now follows descriptions of 10 papers on how machine learning has been used in financial forecasting.

2.1.1 Predicting the direction of stock market prices using random forest [1]

In 'Predicting the direction of stock market prices using random forest', Khaidem, Saha and Dey used random forest to predict the medium and long-term returns of stocks.

1. Input and output

   Before extracting the input features, the author first smoothed the stock price through the exponential moving average. The input features of the model are technical indicators calculated based on processed stock prices, including Relative Strength Index, Stochastic Oscillator, Williams $\%R$, Moving Average Convergence Divergence, Price Rate of Change, On Balance Volume.

   The output of the model is the classification of stocks (two classifications) based on the level of return. This research carried out classification forecasts for stock returns after 30 days, 60 days, and 90 days.

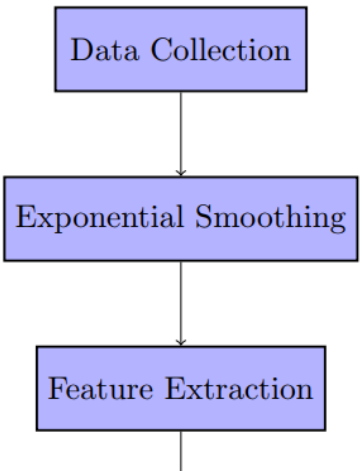


Fig. 2.1 Feature extraction of 'Predicting the direction of stock market prices using random forest' [1]

2. Model

   Random forest classification.

3. Evaluation

   The research used multiple indicators to evaluate the result, including out-of-bag error (OOB rate), Accuracy, Precision, Recall, Specificity, Receiver Operating Characteristic curve (ROC).

4. Result and conclusion

   After applying the model to the three stocks: GE, AAPL and Samsung, the model has achieved good prediction results with an accuracy rate of 85%-95%. At the same time, the 90-day forecast is better than the 60-day forecast, the prediction result of 60-day is better than that of 30-day.

5. Review

   (a) Before feeding the training data to the random forest classifier, this article conducted a linear inseparability test on the binary dependent variable data to verify the necessity of applying a nonlinear model (e.g., the random forest).

   I do not perform a linear inseparability test on the dependent variable data in my research. My opinion is that stock returns are linear inseparable from independent variables. If stock returns are linearly separable, the forecast of stock returns will become too easy.

(b) This article uses a specific python package. If running efficiency is not considered, this python package can theoretically track and explain the meaning of each tree from the highest to the lowest branch.

(c) The most significant potential weakness of this research is that the out-of-bag data error rate may not be a substitute for a real rolling window test. Over time, there are inherent systematic changes in the stock's return and various input features hence the parameters applicable on the training set may not be applicable to the test set. Optimising on the data set only according to OOB can easily lead to over-fitting, which makes the model invalid when used for future prediction. In Section 2.5, I will verify this with experimental results in detail.

2.1.2 Research on the trading strategy of Shanghai and Shenzhen 300 stock index futures based on XGBoost [2]

1. Input and output

   Features are 46 technical indicators calculated based on the opening price, closing price, highest price, lowest price and trading volume of Shanghai and Shenzhen 300 stock index futures. Labels are up and down direction forecast. Researchers use the last five years' data as training set and the last year's data as testing set.

2. Model

   Extreme Gradient Boosting (XGboost).

3. Evaluation

   ROC, confusion matrix

4. Result and conclusion

   The author made predictions for all trading days of stock index futures in 2016, 2017 and 2018. The prediction accuracy of XGboost was 54.66%, and the prediction accuracy of Random Forest was 52.73%.

5. Review

   (a) Both random forest and XGboost algorithms are constructed based on decision tree. The difference is that random forest is an algorithm based on the bagging principle, and XGboost is an algorithm based on the boosting principle.

(b) The potential problem with this paper is that the design of the methodology can only test the classification ability of XGboost on historical data, which can not be applied to make actual predictions for the following reason:

Assuming that the current day is 1st January 2019, according to the methodology in the article, we can use the data from 2013 to 2017 as the training set and the data from 2018 as the prediction set (assuming there are 250 trading days in 2018, we will have 250 binary data for the algorithm to classify) and get a classification accuracy rate. So will stocks rise or fall on 2nd January 2019? The algorithm in the article cannot answer this question because there is not enough binary data to input to the algorithm classification. We will solve this weakness in section 2.5 and give out the answer.

2.1.3 Research on Shanghai and Shenzhen 300 Index trend forecast based on machine learning [3]

1. Input and output

   19 technical indicators: Average Directional Index (ADX), Absolute Price Oscillator (APO), Average True Range (ATR), HT-TRENDLINE, Balance of Power (BOP), Commodity Channel Index (CCI), etc. Three fundamental indicators: Shanghai and Shenzhen 300 Index and IF main contract basis, dynamic PE, dynamic PB.

   Classification results: trend movement and oscillating movement are the input labels and final output.

   Data from January 1, 2015, to November 30, 2017, is the training set. Data from December 1, 2017, to March 31, 2018, is the testing set.

2. Model

   XGboost and Random Forest

3. Evaluation

   Accuracy

4. Result and conclusion

   The out-of-sample prediction accuracy of random forest is 71.25%, and the accuracy of XGboost is 73.75%.

5. Review

(a) One of the most prominent characteristics of this article is that the classification tags are trend movement and oscillating movement. First, through empirical mode decomposition, the closing price in the sample is decomposed. After that, each day's trend component and oscillating component are extracted, and the volatility energy ratio that measures the trend movement and the oscillating movement is calculated.

The oscillating component represents the random walk of the index price. It shows the internal balance characteristics of the market. In contrast, the trend component represents the decisive behavior of investors-made momentum characteristics of the market, in other words, the external imbalance characteristics. When the volatility energy ratio is relatively small, the randomness of the index is high, and the market mainly presents a turbulence pattern. The market trend is significant when the volatility energy is relatively large.

By comparing the frequency distribution of the volatility energy ratio from January 1, 2015, to March 31, 2018, the author selects the threshold value of the volatility energy ratio to determine whether it is a trend label or an oscillating label. It is a major flaw of this article. When determining the volatility energy ratio, the author uses 'future data,' and the data set has already covered out-of-sample data, causing the out-of-sample data to become in-sample data.

(b) At the same time, in the construction of out-of-sample data, this article adopts the same approach literature [2]: historical time series data is used to construct the input matrix, resulting in the method's lack of actual forecasting capabilities.

2.1.4 Integrated long-term stock selection models based on feature selection and machine learning algorithms for China stock market [4]

1. Input and output

The input features are 60 indicators from both financial aspects (Valuation factors, Growth factors, Financial quality factors, Leverage factors, Size factors, Liquidity factors) and the technical aspect (Momentum factors, Volatility Factors, Turnover factors, Technical factors).

The output is classification labels. For the purpose of eliminating the impact of market trends and data noise, stock returns are sorted in descending order every month and then classify the top 30% of the stock as 1 and the last 30% as -1.

2. Feature engineering

   (a) The features may have extreme values, affecting the model and leading to abnormal results. The research applies the following methods to solve the problem:

$$x_{i,new} = \begin{cases} x_m + n \times D_{MAD} & x_i \geq x_m + n \times D_{MAD} \\ x_m - n \times D_{MAD} & x_i \leq x_m + n \times D_{MAD} \\ x_i & \text{else} \end{cases}$$

   $x_{i,new}$ is the processed value, and x_i is the value of the ith variable. x_m is the median of the sequence. D_{MAD} is the median of a sequence of $|x_i - x_m|$, and n is used to control the amplitude of the upper and lower limits.

   (b) Standardize different magnitude of features by the general routine method. Authors adopt Recursive Feature Elimination (RFE) to feature selection. The RFE applied in machine learning will perform multiple training rounds and eliminate the feature with the lowest importance in each round of training.

3. Model

   SVMs, RFs, (Artificial Neural Network) ANN. For each model, the parameters are not fully optimised systematically. The author only chooses the final parameters by comparing the results of a limited combination of several parameters.

4. Evaluation

   (a) The authors used a rolling window method on the data set. Accuracy, AUC(Area under the ROC curve) are applied to evaluate the result

   (b) In addition to the above general evaluation methods of machine learning results, the author has established a simulated trading system to comprehensively compare the forecast results from several indicators such as annualized return, winning rate, sharpe ratio, and maximum drawdown.

   (c) According to the probability of each stock being positively classified from high to low, the stocks are divided into ten groups, then it is possible to analyse whether there is a certain linear correlation between the simulated trading results of each group of stocks and the groups (the author did not specify how the probability is calculated, but since there are several indicators used in this research that change daily, we can infer that this probability should be calculated based on the results of the daily classification).

5. Result and conclusion

   (a) The empirical results of this study show that when RFs is applied to feature selection and stock price prediction simultaneously, the model can obtain the best prediction results.

   (b) By selecting different stock numbers for robustness test, the research uses RF-RFs (random forest as both feature selection and stock price prediction model) to select the stocks with top 1% prediction results and construct a long-short portfolio from 2011 to 2018. The portfolio's annualized return is 29.51%, and the maximum draw-down is only 13.58%.

6. Review

   (a) The author performs features reduction by running the algorithm in a loop and calculating the importance of each feature instead of using general methods such as Principal Component Analysis (PCA) and clustering.

   (b) The problem of extreme values in the features is a vital issue in this type of research. In my research, we can see that the data at both tails of the prediction results are the most unstable, likely to be the impact of extreme values.

   (c) The author forcibly reduced the extreme value of the features to a range built based on the median value. Although this can reduce the influence of extreme values on the algorithm's prediction, this method cannot be used in the financial industry when it comes to the real world because this is using future data.

   (d) The author constructed a portfolio containing long and short positions based on the RF-RFs classification results. This research focuses on the Chinese stock market. Only 50 SSE constituent stocks and 90 Shenzhen Component Index constituent stocks can be shorted theoretically. Even shorting these stocks need to be carried out through a procedure that is not as quickly as in the US.

2.1.5 A machine learning framework for stock selection [5]

1. Input and output

   Features: 244 technical and fundamental features

   Labels: rank stocks according to the return-to-volatility ratio and label the top and bottom stocks as positive and negative, respectively. The middle samples are discarded.

2. Feature engineering

   GA is conducted to optimise the feature selection and select a 114 subset features from a 244 set. The fitness function is the AUC rate on liner regression (LR).

3. Model

   Liner Regression (LR), RFs, DNN (Deep Neural Networks)

4. Evaluation

   Statistical aspect: AUC, Accuracy, Precision, Recall, TPR (True Positive Rate), FPR (False Positive Rate)

   Financial aspect: The researcher established a simulated trading system to compare the forecast results.

5. Conclusion

   (a) Whether before or after feature selection, stacking of DNN and RFs have the best effectiveness compared to other algorithms, which proved the potential of ensemble learning in the financial market.

   (b) There is almost no change in the evaluation scores before and after feature selection, proving that some features do not play a role in the classification process before feature selection.

   (c) The evaluation matrix shows that the recall score of LR and DNN is significantly higher than the precision score, which indicates that LR and DNN are radical models. The recall score of RFs is comparable to its precision score, which indicates that RFs is more likely to be a risk-neutral predictive model. The risk level of stacking of DNN and RFs is between RFs and DNN.

   (d) The stock selection strategy can construct profitable portfolios with returns above the market average.

6. Review

   (a) For conclusion 'All the statistical indexes remain almost unchanged before and after feature selection, which shows some features are indeed redundant'. If the indexes remain unchanged, then the feature selection can only improve this research's efficiency but not effectiveness.

   (b) The research use GA to process feature selection, which give me the enlightenment of a new way to select features.

2.1.6 Predicting stock prices using data mining techniques [6]

1. Input and output

   In the beginning, the data contained 9 features, and this number was reduced manually to 6 features as the other features were found not essential and not having a direct effect. Class labels are the investors' action whether to buy or sell.

Table 1: Attribute Description

Attribute	Description	Possible Values
Previous	Previous day close price of the stock	Positive, Negative, Equal
Open	Current day open price of the stock	Positive, Negative, Equal
Min	Current day minimum price of the stock	Positive, Negative, Equal
Max	Current day maximum price of the stock	Positive, Negative, Equal
Last	Current day close price of the stock	Positive, Negative, Equal
Action	The action taken by the investor on this stock	Buy, Sell

Fig. 2.2 Input features of 'Predicting stock prices using data mining techniques' [6]

   Data source: Arab Bank, Code: ARBK. United Arab Investors Company, code: UAIC. Middle East Complex for Engineering, Electronics and Heavy Industries, code: MECE. The period selected is from April 2005 to May 2007

2. Feature engineering

   The stock data collected are all numerical values. In order to make the data discrete so that the classification model can be labeled, the author uses the following method to carry out the data conversion:

   If the open price, min price, max price, and last price of the stock are greater than the settlement value of the previous transaction, the label is positive. Otherwise, the label is negative.

3. Model

   C4.5 Decision Tree, ID3

4. Evaluation

   Accuracy

5. Result and conclusion

   The resultant classification accuracy of the decision tree model is not very high (around 50% for both algorithms on three companies)

2.1.7 Stock selection with random forest, an exploitation of excess return in the Chinese stock market [7]

1. Input and output

   For each training period, the author generates the input and output as follows.

   Input: For the model with fundamental/technical feature space, the input is a $u * v$ matrix, where u is the sample number and is calculated as the total number of stocks multiplied by the number of trading days in the training period, and v is the number of features.

   Fundamental features: Earnings/Price (EP), Book Value/Price (BP), Sales/Price (SP), Net profits year-on-year, Business income year-on-year, ROA, ROE, market cap. Technical features: 27 technical features

   The author equally split all stocks ranked with excess returns (the difference between the stock return and the Chinese Shanghai Shenzhen Index (CSI index) return in descending order into classes, which are the outputs of the training model.

2. Model

   Random Forest

3. Evaluation

   (a) The author constructed a portfolio based on the classification results. The annual return, Maximal drawdown, Sharpe ratio, Sortino ratio, and Calmar ratio are used to evaluate the portfolio's performance.

   (b) OOB is used to evaluate the accuracy of the classification.

4. Result and conclusion

   (a) The research analysed the dependence of strategy performance on model parameters. The result shows that the number of trees, the number of samples, the training period and the rolling period can affect the classification accuracy.

   (b) The author calculates and analyses different features' importance. Market capitalization is the most prominent factor. Three fundamental factors: EP, BP and SP have a relatively higher importance in determining the stock trend direction.

   Regarding the technical factors, the author concludes that long-term price volume features would mainly account for the long-term excess return.

(c) The research selected different features to construct two different feature sets: multi-feature space and momentum feature space (based on other literature), found that momentum feature space has a better performance.

5. Review

   (a) The research uses the OOB rate as accuracy, which is not convincible as we declared.

   (b) There is a significant difference between the results of multi-feature space and momentum feature space, which would be potential improvement enlightenment for this research.

   (c) Some literature focus on the dependence of strategy performance on model parameters, which is valuable.

2.1.8 Stock market prediction using data mining techniques [8]

1. Input and output

   Four columns of simple data: Open price, close price, highest price, lowest price, on stock Axis Bank, Yes Bank, Central Bank, SBI, ICICI & HDFC Bank.

   Labels: Stock with higher return vs lower return label for KNN classification, return value for support vector regression.

2. Model

   KNN, SVR

3. Evaluation

   Accuracy, Figure plot

4. Result and conclusion

   (a) KNN is applied to the data for 5 years. The accuracy of the test data is around 65%-70%. If the data set is not largely skewed, the accuracy is around 48%-53%.

   (b) Through analysing the figure plot of the result, the author thinks that the SVR result could effectively predict stock price.

5. Review

   Many pieces of research regress second-day stock prices based on the last day's stock price. An algorithm that only uses yesterday's price as an input feature to predict

today's price is not desirable. Because according to the random walk theory, once the midpoint of yesterday's price movement is given, today's prediction result will follow Brownian motion and plotting this result looks like an excellent fitting result. However, it does not have any application significance.

2.1.9 Stock selection with random forest in Chinese market [9]

1. Input and output Financial features:

   The ratio of earnings to price (EP), The ratio of book to price (BP), The ratio of sales to price (SP) , Nets Profits YOY (the growth rate of net profits year on year), Business income YOY (The growth rate of business income year on year), ROA (the return on assets), ROE(The return on equity), Market cap (Market capitalization calculated as price times shares outstanding).

   Technical features are showed in figure 2.3. Labels: 20 day stock return.

Technical features.

Factors	Description	Formula
turnover_20, turnover_40, turnover_60, turnover_120, turnover_240	Refers to the moving average of the turnover over a certain period	$movavg(turnover, m) m \in \{20, 40, 60, 120, 240\}$
close_0/close_9, close_0/close_19, close_0/close_39, close_0/close_59, close_0/close_119	Refers to the momentum with different time lags and can be used to help identify the trend of the price process	$\frac{P_t}{P_{t-m}} m \in \{9, 19, 39, 59, 119\}$
close_19/close_0, close_39/close_0, close_59/close_0, close_119/close_0	Refers to the reversal of momentum	$\frac{P_{t-m}}{P_t} m \in \{19, 39, 59, 119\}$
adjusted_close_0/close_59, adjusted_close_0/close_119	Refers to the momentum with different time lags, excluding the most recent month	$\frac{P_{t-19}}{P_{t-m}} m \in \{59, 119\}$
vol10/vol20, vol10/vol40, vol10/vol60, vol20/vol40, vol20/vol60, vol40/vol60	Refers to a rate of acceleration of a stock's volume and can be used to help identify trend lines of volume	$\frac{movavg(volume, m_1)}{movavg(volume, m_2)} m_1 \in \{10, 10, 10, 20, 20, 40\}\ m_2 \in \{20, 40, 60, 40, 60, 60\}$
volatility_10, volatility_20, volatility_40, volatility_60, volatility_120	Refers to the volatility over the past m trading days as calculated by the standard deviation of daily returns	$movstd(daily.R, m) m \in \{10, 20, 40, 60, 120\}$
std(volume_10), std(volume_20), std(volume_40), std(volume_60), std(volume_120)	Refers to the standard deviation of trading volume time series over the past m trading days	$movstd(volume, m) m \in \{10, 20, 40, 60, 120\}$

Fig. 2.3 Technical features of Stock selection with random forest: An exploitation of excess return in the Chinese stock market [9]

2. Model

   Random Forest Classification

3. Evaluation

   The researchers constructed a portfolio with stock in the 'upper class', where the first 20 stocks with the highest probabilities are selected and held for a certain period until the next stock ranking date. The benchmark is CSI 500 index, and the transaction cost is 0.16%

4. Result and conclusion

   The critical contribution of this research is that the researchers tested the extent that the strategy's performance depending on the RFs model hyper-parameters, including the number of trees, the number of sample classes, the training period and the rolling period.

   When the tree number is set to 60, the Sharpe ratio can reach 2.75, and the Sortino and Calmar ratios also reach the maximum. However, it should be noted that more trees increase the OOB accuracy does not mean that it can increase the out of sample accuracy.

   With the increase in the number of sample classes, the out-of-sample performance of the portfolio deteriorated, especially in 2016 and 2017. The increase in sample categories led to a decrease in excess returns. The change of training period and the rolling period has an almost indistinguishable impact on the result.

2.1.10 Impact of financial ratios and technical analysis on stock price prediction using random forest [10]

1. Input and output

   63 features, including quarterly financial ratios and technical ratios consisting of 433 companies listed in the Hong Kong Stock Exchange from 2011-2014. Labels are quarterly stock returns.

2. Model

   Random Forest Classification

3. Evaluation

   Accuracy

4. Result and conclusion

   The results show that the result is just slightly better than random results. This indicated that using financial ratios to predict the next quarter's results were unreliable.

2.2 Literature Review on Linear Model Prediction

2.2.1 Twitter mood predicts the stock market [11]

In this paper, the authors' purpose is to find the connection between 'Twitter Mood' and 'Stock Market.' The article applies two models: linear Granger causality test and machine learning,

1. Input and output

   The input mood features are extracted by opinion finders and Google Profile of Mood States (G-POMS), the input and out put labels are Dow Jones Industrial Average (DJIA). The 'mood' of Twitter users was quantified through the opinion finder and G-POMS, and eight mood features were extracted. DJIA was used to represent the stock market index as labels, and these tools made quantitative research possible. Before put features into the model, some features engineering technique was applied: normalization, features combination (combine different mood type for self-organizing fuzzy neural networks).

2. Model

   Granger causality & self-organizing fuzzy neural networks (SOFNN). The Granger causality test was used to verify the causal relationship between mood and the stock market. By adding mood features as input to a price prediction model (SOFNN) and comparing the results before and after adding, the causal relationship between mood and stock market can be cross-validated.

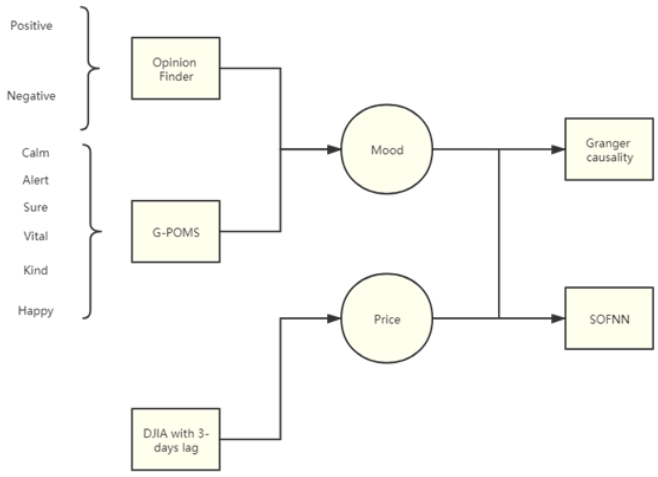


Fig. 2.4 Methodology structure of 'Twitter mood predicts the stock market'

3. Evaluation

   F statistic, P-value (for Granger causality), Mean Absolute Percentage Error (MAPE) and direction judgement accuracy (for SOFNN)

4. Result and conclusion

   Only the 'Calm' feature is valuable, and it usually has a 3-4 days shift when reflected on DJIA. Combined with 'Calm,' 'Happy' could also increase the accuracy of the SOFNN direction judgment, while 'Happy' does not even pass the Granger causality.

2.2.2 Fama French three-factor model and five-factor model [12, 13]

Besides non-linear models such as machine learning, scholars widely use linear models to predict stock returns. The Fama French three-factor model is one of the most well-known linear models for predicting stock prices.

The three-factor model is an extension of the CAPM model. The CAPM model is built on the following assumption:

1. There is a positive linear relationship between the expected return of a security asset and its Beta (the price volatility of a stock relative to the entire stock market).

2. Beta is sufficient to explain the expected return of a security asset.

However, some scholars have found that Beta cannot fully explain the excess return of an asset. Empirical studies have shown that stock market value, book-to-market ratio, financial leverage, and price-earnings ratio have explanatory effectiveness on stock excess returns. At the same time, CAPM cannot explain these anomalies.

Fama and French studied the relationship between the average returns of stocks (except financial stocks) traded on NYSE, AMEX, and NASDAQ from 1963 to 1990 and these factors. After performing multiple regressions, they found that the size factor and book-to-market factor, plus the original Beta, can consistently explain the average return of stocks.

Fama and French published the five-factor model in 2013. The five-factor model has two more factors than the three-factor model: earning factor is the difference between the returns of high/low earnings stocks, reinvestment factor is the difference between the returns of low/high reinvestment stocks. These two factors respectively measure the level of profitability risk and the level of reinvestment risk (the ability of the company to expand reproduction). Similar to the three-factor model, the method for parameter estimation of the five-factor model is still multiple linear regression. After empirical testing, the author believes that the five-factor model has better explanatory power than the three-factor model.

2.3 Supporting Literature

2.3.1 Do we need hundreds of classifiers to solve real world classification problems? [14]

When researchers choose a classification/regression model, they are limited by their conditioned background within computer science and mathematics. For example, some classifiers (linear discriminant analysis or generalized linear models) come from statistics, while others come from symbolic artificial intelligence and data mining.

In this research, authors evaluate 179 classifiers arising from 17 families (discriminant analysis, bayesian, neural networks, support vector machines, decision trees, rule-based classifiers, boosting, bagging, stacking, random forests and other ensembles, generalized linear models, nearest-neighbours, partial least squares and principal component regression, logistic and multinomial regression, multiple adaptive regression splines and other methods).

The research is implemented in Weka, R, C and Matlab. They used 121 data sets from the whole UCI database and other real data sets.

The classifiers most likely to be the bests are the random forest (RFs), the best of which (implemented in R and accessed via caret) achieves 94.1% of the maximum accuracy and achieves overcoming 90% in 84.3% of the data sets. The difference is not statistically significant with the second-best, the SVMs with Gaussian kernel implemented in C using LibSVMs, which achieves 92.3% of the maximum accuracy.

A few models are better than the remaining ones: random forest, SVMs with Gaussian and polynomial kernels, extreme learning machine with Gaussian kernel, C5.0 and avNNet (a committee of multi-layer perceptrons implemented in R with the caret package). The random forest is the best family of classifiers (3 out of 5 bests classifiers are RFs), followed by SVMs (4 classifiers in the top-10), neural networks and boosting ensembles.

2.3.2 Comparison of two exploratory data analysis methods for fMRI: Unsupervised Clustering Versus Independent Component Analysis [15]

The authors conducted a detailed comparative research among unsupervised clustering methods: 'neural gas' network, fuzzy clustering, Kohonen's self-organizing map, spatial Independent Component Analysis (ICA), topographic ICA, PCA. All the methods are tested on the fMRI data set.

The clustering results were evaluated by a. task-related activation maps. b. associated time-courses. c. receiver operating characteristic (ROC) analysis.

The clustering methods outperform the transformation-based methods for all subjects from the evaluation analysis. Both the 'neural gas' network and fuzzy clustering based on deterministic annealing outperform ICA in terms of classification results but require a longer processing time than the ICA methods.

2.3.3 Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs. principle component analysis [16]

The authors tested the performance of two data-driven methods (FCA, PCA) when applied to Functional magnetic resonance imaging (fMRI) data analysis.

In a routine simulated test, two types of data with different noise characteristics obtained from an magnetic resonance (MR) experiment were used: a. Water phantom data with scanner noise only b. MR time series acquired under the null condition with scanner and physiological noise present.

The results suggest that

1. If the time series is corrupted with scanner noise only, both methods show comparable performance.

2. If other sources of signal variation (e.g., physiological noise) are present, PCA fails to identify activation, which may be critical in fMRI. FCA outperforms PCA in this condition.

Due to the arbitrary sign of the eigenvectors obtained from the correlation matrix decomposition (PCA's essential calculation process), PCA can not immediately distinguish between positively and negatively correlated time courses. FCA preserves the original formations of the time-courses and, as a consequence, could yield immediately interpretable results.

2.4 Literature Review Summary

There are the following conclusions after summarizing the above literature:

1. RFs and SVMs are broadly used in the security prediction area. Literature [1, 3–5, 7, 9, 10] (described in section 2.1.1, 2.1.3, 2.1.4 , 2.1.5 , 2.1.7 , 2.1.9 and 2.1.10 respectively) use random forest classification as a prediction model, paper [4] in section 2.1.4 uses SVMs classification as a prediction model, and paper [9] in section 2.1.9

uses SVR as a prediction model. Both types of models have achieved compelling results.

2. Referring to the above and further literature results, most studies use classification models as the prediction model (our research will use two regression models: SVR and RFR as the primary model).

3. Judging from the input features, Papers [5, 6, 8, 10] (described in section 2.1.5, 2.1.6, 2.1.8 and 2.1.10 respectively) use both financial and technical indicators, while literature [2] in section 2.1.1 and 2.1.2 only use technical indicators, and both have achieved compelling predictions. However, it is worth noting that literature [7, 9] in section 2.1.7 and 2.1.9 use too simple technical indicators, and the prediction result is not significant.

4. Both FCA and PCA are effective feature engineering methods. On some indicators and data sets, FCA outperformed PCA [15, 16].

5. A few models perform more effectively on the literature's data set: random forest, SVMs with Gaussian and polynomial kernels, extreme learning machine with Gaussian kernel, C5.0 and avNNet. The random forest is clearly the best family of classifiers, followed by SVMs [14].

2.5 Verification of Appropriate Evaluation Methodology

In the literature review part, we have reviewed works of literature that ignores the difference between time series feature data (including time series data and panel data) and cross-sectional data, resulting in two potential mistakes:

Mistake 1: Some papers equate the OOB accuracy with the actual testing accuracy.

The random forest is established on the bagging principle (see Section 6.4 for details). In bagging, it can be found that about 1/3 of the samples of the bootstrap method will not appear in the sample set, so they did not participate in the establishment of the decision tree. These data are called out-of-bag data.

Therefore, some literature believes that verifying the model on out-of-bag data is equivalent to verifying the test set. The out-of-bag error rate of the training set can be used to replace the error rate estimation method of the test set.

Assuming that there is a period sequence data or panel data, we divide it into the training set and test set, and the OOB rate generated by bootstrap will be generated from the training

set. Assuming that the data structure does not change over time, this method is feasible, but unfortunately, time-series data usually varies in different periods.

Mistake 2: As we reviewed in literature [1–3] in Section 2.1.1, 2.1.2 and 2.1.3, the methodology of many papers divide historical data into the training set and testing set in a time-series manner, for example, Stock X, Training set: data of 1 to 200 days, Test set: data 200 to 300 days.

In the following part of this section, we will reproduce the experiment in Section 2.1.1 and verify the above two mistakes.

2.5.1 Data and methodology

According to paper 'Predicting the direction of stock market prices using random forest' [1] in Section 2.1.1, we replicate entirely their methodology construct the same random forest classification program with crucial issues:

1. The input features of the model are technical indicators calculated based on processed stock prices, including Relative strength index, Stochastic Oscillator, Williams %R, Moving Average Convergence Divergence, Price Rate of Change, On Balance Volume.

2. The input labels are 1 and 0, representing stock with high return and stock with low return correspondingly.

3. The hyper-parameters remain the same. For detailed methodology, please refer to page 3 to page 13 of the article 'Predicting the direction of stock market prices using random forest' [1].

The only difference between this verification and their research is the data. They use AAPL, GE dataset (Listed on NASDAQ) and Samsung Electronics Co. Ltd. (Listed on Korean Stock Exchange). In our verification, the data is: Stock code: 000002 WANKE from Shenzhen Stock market exchange, Data period: 1st Jan 2018 - 1st Aug 2020. The stock price chart of the period is shown in Figure 2.5. From the figure we can see that the start and the end of the period are nearly on the same price.

2.5.2 Result: OOB accuracy vs real test accuracy

The paper [1] has three main findings and our replication achieve all of them:

1. The paper uses the OOB error rate as the testing error rate and concludes that the OOB error rate of 60 days is 15%. The result is shown in the left picture of Figure 2.6.

Fig. 2.5 Price of 000002 WANKE, Shenzhen Stock Market Exchange, 1st Jan 2018 - 1st Aug 2020

Our replicated model achieves 17% of 60 days OOB error rate, which is shown in the right picture of Figure 2.6.

2. The paper declares that 90-day forecast is better than the 60-day forecast, the prediction result of 60-day is better than that of 30-day.

   In contrast, we can declare same conclusion based on the result.

3. In their paper, they verify that the OOB rate gradually decreases with the increase in the number of trees.

   In contrast, we observed same decreasing process in the right picture of Figure 2.6.In the end, the OOB rate of both programs has converged below 15%.

From the above contrast, results based on the same methodology are very close. If we use the same evaluation method, we will come to the same conclusion.

However, in order to verify whether OOB accuracy (1 minus OOB error) can be equal to actual testing accuracy, we applies real testing data instead of out of bag data.We split the data into the training set and test set at a ratio of 6:4. We run the program on the test set, the accuracy obtained from the testing set is 56%. This real test accuracy is far below OOB accuracy, which means the OOB accuracy can not be used as test accuracy.

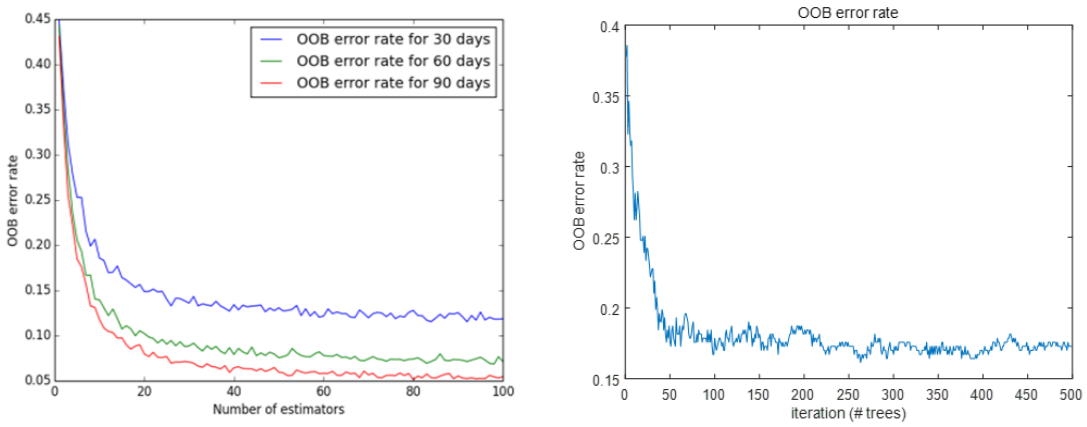


Fig. 2.6 OOB error rate of the contrast paper and verification

2.5.3 Result: two dividing data method

In Section 2.5.2, we proved that OOB accuracy is not equal to test accuracy. Although the test accuracy we obtained is theoretically correct, this method cannot be used to guide predictions. As discussed, we split the time series historical data into training and testing sets to get accuracy. However, we cannot know the classification result of tomorrow because we cannot make a classification on one data.

We can apply the following data set construction method to solve the above problem:

Assume the price of tomorrow is the price we want to classify. We can set a certain range of past days plus tomorrow as as the test set so that we can get the classification result of tomorrow. Every day, we include the data of a new day in the test set and exclude the data of the first day, so this is a rolling window method with a step of 1 day. After running in this way for days, we can get the prediction result for every day, and in contrast with the actual data, we can calculate the accuracy. This is a guiding accuracy for actual investment, and for convenience, we call this method 2, and the method in Section 2.5.2 is method 1.

We apply this method to the same training and testing data set in Section 2.5.2, and the prediction accuracy is 0.513. In order to enhance the rigor of the result, we use two dividing data methods (method 1 and method 2) to test all the stocks of the Shanghai and Shenzhen stock exchange of the same period. The corresponding accuracy of the two methods is 0.52 and 0.49.

The above accuracy is not significant.

2.5.4 Three types of data

From the analysis in subsection 2.5.2 and subsection 2.5.3, we can conclude that:

1. No matter which data construction method we adopt or whether the data construction method can be applied to actual transactions,OOB accuracy cannot replace the actual test accuracy.

2. For market time series data, combining data from time axis into same data set (no matter training or testing) may be not effective

The above analysis remind us to pay attention to the difference in data sets when applying machine learning. The data sets have three types: time series, cross-sectional, and panel data.

1. Cross-sectional data

   Cross-sectional data refers to the data of different objects collected at a certain time. It corresponds to a one-dimensional data set composed of different spaces (objects) at the same point in time. It studies a specific economic phenomenon at a certain time and highlights the differences in space (objects). Usually, cross-sectional data shows irregular rather than random changes, which is the so-called 'heterogeneity' in econometrics.

2. Time series data

   Time series data refers to the data obtained by continuously observing the same object at different times. It focuses on the changes in the time sequence of the research object, looking for the law of the diachronic development of the space (object). When using time series as samples, researchers should pay attention to the consistency of data behaviour in the selected sample interval.

3. Longitudinal data or panel data

   Panel data is a data resource that combines cross-sectional data and time series. It can be used to analyse the characteristics of the data composed of each sample in the time series.

In summary, both time series data and panel data have time-series characteristics: with the development of time, the behaviour of the sample interval data may be inconsistent. I think this is the fundamental reason for the distortion of OOB rate in some papers: OOB rate essentially uses training data that is not used by random forest to instead real data. This part of out of bag data and in bag data are in the same time period, while the real test data is in another time period.

When using real test data, the time-series characteristics may caused the failure in Section 2.5.3 for two reasons:

1. The training and testing data is from different period.

2. Each single data from test set is from different days.

In our methodology, we avoid reason 2 by using data from same day as testing data set. It is essentially cross-sectional data cut from panel data. Further details could be found in Section 7.2. We fixed the problem caused by reason 1 by mathematical technique in Section 7.4 and subsection 7.5.2.

2.6 Conclusion

In the literature review, we reviewed the representative literature related to this research and conducted the experimental test on some conclusions in the literature, from which we can get the following enlightenment:

1. Using simple technical indicators as input to predict the results is insufficient. Combining financial and technical indicators to build a multiple feature space is better.

2. RFs and SVMs are effective classification models.

3. Both FCA and PCA are effective feature reduction methods, and FCA has some advantages over PCA.

4. OOB accuracy cannot directly replace actual accuracy for time series forecasting.

5. We need to find effective method to avoid the distortion caused by data with time-series characteristics.

In summary, our research adopts FCA as a feature reduction method, RFs and SVMs as prediction model, and financial indicators and technical indicators as the feature input, which would be a feasible technical path.

Chapter 3

Feature Extraction

3.1 Introduction

As we showed in Figure 1.4 and 1.5, before running a machine learning algorithm, we need to construct input feature for the model. Our support vector regression and random forest regression share the same input feature. In this chapter, we will introduce the process of feature extraction.

In the first version of the research only financial ratios are used as input features. Strictly audited financial indicator data is disclosed in annual reports, and only using annual report data will cause the problem of a low frequency of input data fluctuations. We can solve this problem by adding quarterly report data, but there are three problems in implementation:

1. The quarterly report is only revised internally by the company and therefore has has not been audited. Furthermore, listed companies have the motivation to modify their performance. Untrue indicator data will have a negative impact on our predictions.

2. The quarterly report data is incomplete. According to Article 12 of the China Securities Regulatory Commission's 'Administrative Measures on Information Disclosure of Listed Companies': For the periodic reports of listed companies, only annual reports and interim reports must be compulsorily disclosed. Even if the company chooses to disclose the quarterly report, some financial data will be missing compared to the annual report. Missing data adversely affects our prediction.

3. On one hand, differences between firms in a pure accounting ratio such as ROE can be expected to be already reflected in the share prices, and if so, they may not have predictive power. On the other hand, financial ratios differ from industrial sector to sector. Hence, the difference of ratios from individual stocks may be caused by

different sector averages (e.g., the high-tech industry usually has a higher average PE than heavy industry). The above two aspects may deteriorate the predictive power of financial ratios.

Based on the above reasons, in addition to selecting quarterly report (as long as the firm disclosed) financial data as the model input, we added the technical indicators that change daily as the model input.

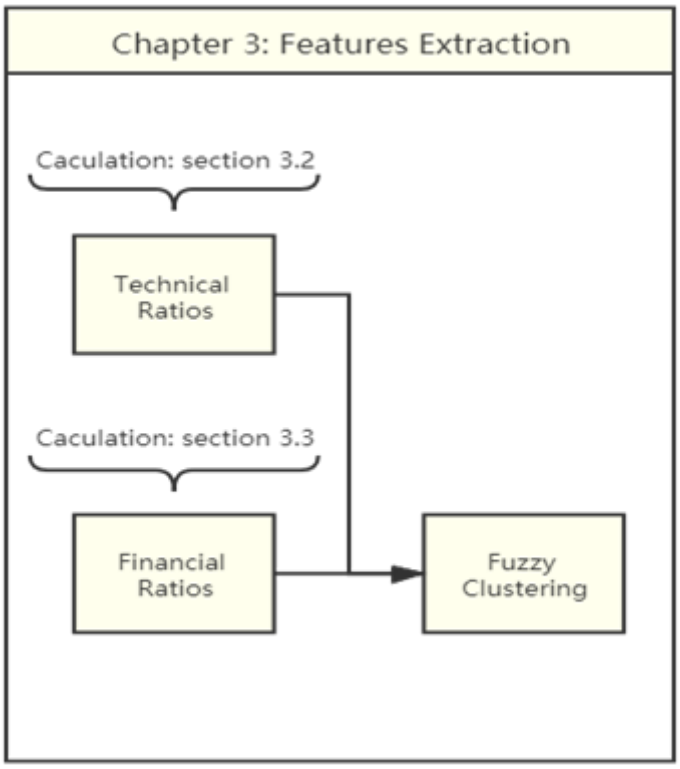


Fig. 3.1 Feature extraction pipeline

3.2 Technical Feature Calculation

In this section, we calculated 8 technical indicators listed below. It must be noticed that MACD & MACD signal contains two ratios. Hence 8 technical indicators will generate 9 input features.

1. Long Term Relative Position Index

The formula for calculating Long Term Relative Position Index is

$$OI = 2 * C - L130 - H130$$

where

$C =$ Closing Price on the day

$L130 =$ Lowest Low Price over the past 130 days

$H130 =$ Highest High Price over the past 130 days

The long-term relative position index measures the position of the stock's current price in a long time interval (130 days). If the index is large, it indicates the relative high of the stock price in the range, and if the index is small, it indicates the relative low of the stock price in the range.

2. Trading Volume

   Trading volume is the number of securities transactions in a specific period directly given by the exchange. Trading volume is an important indicator. When special events do not drive the market or individual stocks, the trading volume is a random function and has nothing to do with the price. Therefore, sudden changes in trading volume are almost necessarily accompanied by drastic changes in stock prices. In the algorithm, we use 1-day trading volume as an input feature.

3. Relative Strength Index

   The formula for calculating RSI is:

$$RSI = 100 - \frac{100}{(1 + RS)}$$

$$RS = \frac{AverageGainOverpast14days}{AverageLossOverpast14days}$$

   RSI is a popular momentum indicator among investors. It can determine whether a stock is overbought or oversold. When demand pushes the price to a high position for no apparent reason, the stock is overbought. This situation is usually interpreted as the stock being overvalued, and the price may fall. When the price drops sharply below its instinct value, the stock is said to be oversold, and this is the result of panic selling. The RSI ranges from 0 to 100. Generally, when the RSI is higher than 70, it may indicate that the stock is overbought. When the RSI is lower than 30, it may indicate that the stock is oversold.

4. Stochastic oscillator

$$\%K = 100 * \frac{C - L14}{H14 - L14}$$

where

$C =$ Closing Price on the day

$L14 =$ Lowest Low Price over the past 14 days

$H14 =$ Highest High Price over the past 14 days

It can be seen from the formula that the Stochastic Oscillator (also known as KD, %K) measures the closing price level relative to the low and high ranges within 14 days. In essence, the stochastic oscillator is similar to the RSI. Both indicators measure the overbought and oversold levels of stocks. When the stochastic oscillator is too high, the stock price is likely to fall, and vice versa.

5. MACD & MACD signal

$$MACD = EMA_{12}(C) - EMA_{26}(C)$$
$$\text{Signal Line } = EMA_9(MACD)$$

where

$MACD =$ Moving Average Convergence Divergence

$C =$ Closing price series

EMA $_n = n$ day Exponential Moving Average

6. PROC

$$PROC(t) = \frac{C(t) - C(t-n)}{C(t-n)}$$

where $PROC(t)$ = Price Rate of Change at time t

$C(t)$ = Closing price at time t

It measures the most recent change in price with respect to the price in n days ago, where $n = 14$.

7. OBV

$$OBV(t) = \begin{cases} OBV(t-1) + Volume(t), if\ C(t) > C(t-1) \\ OBV(t-1) - Volume(t), if\ C(t) < C(t-1) \\ OBV(t-1), if\ C(t) = C(t-1) \end{cases}$$

where

$OBV(t)$ = On Balance Volume at time t

$Volume(t)$ = Trading Volume at time t

$C(t)$ = Closing price at time t

OBV(on balance volume) regards the trading volume when the stock price rises as the accumulation of popularity, while the trading volume on the day when the stock price falls is regarded as the dispersion of the popularity and perform subtraction.

8. Williams %R

$$W\%R = \frac{H14 - C}{H14 - L14} * -100$$

where

C = Closing price on the day

$L14$ = Lowest Low over the past 14 days

$H14$ = Highest High over the past 14 days

The value range of Williams %R is -100 to 0. When its value is greater than -20, a sell signal is generated. When its value is lower than -80, a buy signal is generated.

3.3 Financial Feature Calculation

We divide the company's financial indicators into five categories: profitability indicators, development capability indicators, shareholders' equity indicators, solvency indicators and operating capability indicators. These financial indicators can be obtained through the disclosed financial statements.

1. Profitability

   Profitability refers to the ability of a company to obtain profits, which mainly reflects the relationship between profits, income and assets. A company's profitability is the foundation of its survival and development, and it is also the core factor supporting the stock price. Many financial indicators reflect a company's profitability, such as profit margin, gross profit margin, net profit margin, return on equity.

2. Development capability

The development capability of a company reflects the development prospect of the company in the coming years. It is estimated based on the sustained growth of the company's profitability. Development capability indicators are the basis for securities investors to make the long-term evaluation. Hence, investors who focus on the companies' future vision will pay more attention to these indicators.

3. Shareholders' equity

   The shareholder's equity indicators measure the company's ability to directly give the company's shareholders a return on investment, which is usually equivalent to the company's ability to pay dividends. In the long run, the company's profitability is consistent with the shareholder's equity, but if the company's decision-makers decide not to pay dividends in the short term, then shareholders' equity indicators and profitability indicators will deviate. Such indicators are significant for small and medium investors who cannot influence the company's decision-making and risk-averse investors who expect stable cash flow income.

4. Solvency

   The solvency indicators measure the company's ability to repay debts (including short-term and long-term debt). In order to maintain commercial operations, the enterprise must hold enough cash and cash equivalents to pay various due debts. An enterprise can seek profit and development only based on not going bankrupt. Once the enterprise goes bankrupt, all prospects will become a mirage. Generally speaking, the pressure of enterprises to repay debts mainly from two aspects: First, pay the principal and interest of ordinary debts, such as long-term loans, bonds payable and short-term settlement debts. Second, pay taxes. Solvency indicators are not only the indicators that long-term investors pay attention to but also hostile bidders.

5. Operating capability

   Operating capability refers to the efficiency of the enterprise's use of internal human resources and production materials under the constraints of the external market environment. Operating capability indicators mainly include accounts receivable, inventory, current assets, fixed assets and total asset turnover. These indicators can analyse how managers use management skills to give full play to the operational efficiency of assets under a given scale of assets. In layman's terms, operating capability reflects a company's utilization of the assets. For example, real estate companies A and company B invested 100 million in cash to build real estate. If A built one building and B built two buildings, company B's operating capability would be more robust.

We extracted financial information and calculated financial ratios of all the stocks in Shanghai and Shenzhen Stock Market, China, from 2009 to 2018. The data are cleaned in Excel, extracted and calculated in MATLAB. 42 financial ratios from 5 categories are calculated as the original data before fuzzy clustering.

Each year's financial data are published by the listed company no late than 30th April of the following year. Financial features are recalculated on a quarterly frequency. The demonstrated result example in Section 3.7 is from 2014 (which could be collected before 30th April 2015).

42 financial ratios from 5 categories	
F1 Profitability ratios	F3-8 Book to market ratio (Parent Statement)
F1-1 Operating gross profit margin	F3-9 EPS/P
F1-2 Operating net profit margin	**F4 Solvency ratios**
F1-3 Return on assets before interest and tax	F4-1 Current ratio
F1-4 Return on assets	F4-2 Quick ratio
F1-5 Return on current assets	F4-3 Debt ratio
F1-6 Return on fixed assets	F4-4 Shareholder's equity to liabilities ratio
F1-7 Marginal profit ratio	F4-5 Debt to market value ratio
F1-8 Return on equity	F4-6 Fixed assets to total assets ratio
F1-9 Growth rate of main business income	F4-7 Equity to total assets ratio
F2 Development capability ratios	F4-8 Working capital to total assets ratio
F2-1 Appreciation rate of capital preservation	F4-9 Working capital to net assets ratio
F2-2 Capital accumulation rate	F4-10 Owner's equity ratio
F2-3 Fixed assets growth rate	F5 **Operating capability ratios**
F2-4 Total assets growth rate	F5-1 Accounts receivable Turnover
F2-5 Net profit growth rate	F5-2 Inventory turnover
F3 Shareholders' equity ratios	F5-3 Accounts payable turnover
F3-1 Operating income per share	F5-4 Working capital turnover
F3-2 Net assets per share	F5-5 Current assets turnover
F3-3 Surplus reserve per share	F5-6 Fixed assets turnover
F3-4 Capital reserve per share	F5-7 Long term assets turnover
F3-5 Undistributed profit per share	F5-8 Total assets turnover
F3-6 Market to book ratio	F5-9 Equity turnover
F3-7 P/E ratio	

Table 3.1 Financial ratios from 5 categories

3.4 The Reason of Applying FCA and Non-technical Overview

Through features calculation in Section 3.2 and 3.3, we obtained 42 financial features and 9 technical features. When there are too many feature dimensions or the high coincidence of feature dimensions, the feature dimensions need to be reduced. This process is called feature extraction. There are two ways to reduce the feature dimension. One is feature selection, and the other is feature reconstruction.

Selecting the most compelling features from the feature set to reduce the dimensionality of the feature space is called feature selection. Feature selection directly ignores the features that have no or little contribution to the class separability.

Feature reconstruction is the mapping or conversion of high-dimensional features into low-dimensional features. The feature obtained by feature reconstruction is a specific combination of the original features set. That is, the new feature contains the information of the original feature.

The selection and reconstruction of features are crucial, directly affecting the performance of supervised machine learning classifiers and regressors. If the difference between different features is significant, it will be easier to design a classifier or regressor with higher perform.

In this research, we do not perform data engineering for stock technical indicators. For stock financial indicators, we will apply fuzzy clustering to filter them for the following reasons:

1. 42 financial indicators in five categories exceed the need for input and lower the algorithm's efficiency.

2. The calculation formulas of the indicators in each category share the same accounting statistics, so there may be multicollinearity. For example, liquidity ratios such as current ratio and quick ratio are highly correlated.

3. If we directly filter from these indicators based on experience, it may cause two problems: a. Missing indicators that are genuinely explanatory b. Still have multi-collinearity among the selected indicators.

Therefore, a quantitative method should be used for indicators screening, which reduces the massive amount of information to an operable range while ensuring the integrity of the information.

There are various input feature dimensionality engineering methods, which can be divided into labelled dimensionality engineering method and unlabeled dimensionality engineering method according to the presence or absence of labels.

Two typical unlabeled dimensionality engineering methods are PCA (Principal Component Analysis) and FCA (Fuzzy Cluster Analysis). In this study, we use FCA to deduct dimensionality for three reasons:

1. According to the literature' Comparison of two exploratory data analysis methods for fMRI: fuzzy clustering vs principle component analysis and the literature' Comparison of Two Exploratory Data Analysis Methods for fMRI: Unsupervised Clustering Versus Independent Component Analysis', the test result of FCA on fMRI data set is better than PCA [15, 16].

2. Different classification models have their own optimal data sets, so we cannot conclude in general that FCA will outperform PCA on the data set used in this research. Referring to the literature, it can be found that both PCA and FCA are commonly used feature reduction methods for machine learning. Similar to those research, this research aims not to compare PCA and FCA or dozens of other reduction methods to answer which one has the best effectiveness. At the step of feature reduction, as long as an effective method is selected, the requirements of this research can be met.

3. The principal component analysis will construct a new component based on the correlation between the original features, thereby changing the original feature names (for example, PE ratio and PB ratio, if they are highly correlated, will be combined into a new component), compared to FCA, changing the feature name will result in reduced interpretability.

Clustering is a type of unsupervised learning. In unsupervised learning, the labelled information of training samples is unknown, and the goal is to explain the inherent relation of the data by learning unlabeled training samples.

Clustering divides the data set into subsets, and each subset is called a cluster, which may correspond to some potential class. It should be noted that the corresponding meaning of the cluster is unknown, the clustering process can only automatically form a cluster structure, and meaning corresponding to the cluster needs to be given by the user. Fuzzy clustering has the characteristics of high efficiency and less information loss. It is a commonly used statistical method that can screen large data sets.

3.5 Non - technical Overview of FCA

To understand fuzzy clustering analysis (FCA) from the bottom layer involves a series of interrelated mathematical knowledge such as set theory, fuzzy mathematics and matrix op-

erations, and the different proper nouns used between various discipline branches increase the difficulty for outsiders to understand . This section attempts to give readers an understanding of fuzzy clustering analysis without the use of proper nouns and aforementioned mathematical details.

Clustering is easy to understand: according to certain clustering standards, the elements are allocated to different sets. Different sets are called 'clusters' in clustering theory. Roughly speaking, clustering is classifying objects.

When clustering, if each element is classified into an independent set, it is called hard clustering, and if an element can belong to both set A and set B, it is called soft clustering. In later case, the affiliation of the element and the set is 'fuzzy', hence soft clustering is called fuzzy clustering.

The presence of fuzzy in clustering depends on the clustering criterion. E.g:

Element: Jack

Set A: Good Guys

Set B: Bad Guys

Clustering standard: Whether or not Jack has been in prison.

The above case is called hard clustering since we have a hard criterion. If we set the clustering criteria as: what others say about Jack. Jack may not be liked by everyone, then he may have 0.7 in set A and 0.3 in set B, which is fuzzy clustering.

In this thesis, we cluster time series financial features into clusters by the criterion of 'correlation', and then calculate the correlation between elements within each cluster to select representative features. Since the multi-correlation (In terms of Accounting: multi-relationship) between financial features implies the idea of 'fuzzy' in fuzzy mathematics, FCA is a natural fit for classifying this dataset.

To conclude, this chapter applies the mathematical techniques of FCA to classify and select features according to the 'correlation' between features.

3.6 The Process of Fuzzy Clustering

Set $U = (x_1, x_2, ..., x_n)$ as the feature set need to be classified with n dimensions, and each feature contains m components, quote as $x_i = (x_1, x_2, ..., x_{mi})^T$

The original data matrix is presented as Equation 3.1:

$$X = \begin{vmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{vmatrix} \tag{3.1}$$

The clustering process follow steps:

3.6.1 Data standardization

The features need to be standardized to the same magnitude to process a fuzzy clustering, and the features need to be standardized into the interval [0, 1].

The data standardization includes two steps

1. Standard deviation transformation

$$x'_{ik} = \frac{x_{ik} - \bar{x_k}}{s_k} (i = 1, 2, ..., m, k = 1, 2, ..., n) \tag{3.2}$$

where $\bar{x_k} = \frac{1}{m} \sum_{i=1}^{m} x_{ik}$, $s_k = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_{ik} - \bar{x_k})^2}$ after transformation all the means of variables equal to 0, and the standard deviation equal to 1, which realized the dimensionless of data, however, this step has not transformed all s'_{ik} drop into the interval [0, 1] yet.

2. Range transformation

$$x''_{ik} = \frac{x'_{ik} - \min\{x'_{ik}\}_{1 \le i \le m}}{\max\{x'_{ik}\}_{1 \le i \le m} - \min\{x'_{ik}\}_{1 \le i \le m}} (i = 1, 2, ..., m, k = 1, 2, ..., n) \tag{3.3}$$

Obviously all the data x''_{ik} are in [0, 1] after data standardization

Matlab code:

Algorithm 3.1 Data standardization

```
[m,n]=size(a);
b=mean(a);
c=std(a,1);
for j do=1:n
    r(:,j)=(a(:,j)-b(j))./c(j);
end for
for j do=1:n
    d(:,j)=(r(:,j)-min(r(:,j)))./(max(r(:,j))-min(r(:,j)));
end for
```

3.6.2 Construct fuzzy similarity matrix

Define fuzzy relation: set X, Y as two nonempty sets, then the direct product $X \times Y = \{(x,y)|x \in X, y \in Y\}$ contains a fuzzy subset R , which is the fuzzy relation from X to Y.

The fuzzy relation R can be described by the function $\mu_R : X \times Y \to [0,1]$,the degree of $\mu_R x, y$ is the correlation between (x,y) and fuzzy subset R , marked as $R(x,y)$ when $X = Y$, if fuzzy relation $R = \Im(X \times X)$ satisfy:

$$\text{Reflexivity:} R(x,x) = 1;$$
$$\text{Symmetry:} R(x,y) = R(y,x)$$

Then R is the fuzzy similarity relation on X, $R(x,y)$ describe similarity relation on x and y.

The general method to construct fuzzy similarity matrix contains distance method, correlation method, cosine method [21]. In this thesis we use correlation method.

Set $x_i = (x_{i1}x_{i2},...,x_{ip})^T$ and $x_j = (x_{j1}xj2,...,x_{jp})^T$ as two random variables in the variable space, then the correlation coefficient between them is defined in Equation 3.4.

$$r_{ij} = \frac{\sum_{k=1}^{m}(x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j})}{\sqrt{\sum_{k=1}^{m}(x_{ki} - \bar{x_i})^2} * \sqrt{\sum_{k=1}^{m}(x_{kj} - \bar{x_j})^2}} \tag{3.4}$$

where $\bar{x_i} = \frac{1}{m}\sum_{k=1}^{m} x_{kj}$, $\bar{x_j} = \frac{1}{m}\sum_{k=1}^{m} x_{kj}$,i,j=1,2,...,n.

The fuzzy similarity matrix constructed based on correlation method is:

$$R = \begin{vmatrix} R(x_1,x_1) & R(x_1,x_2) & \dots & R(x_1,x_n) \\ R(x_2,x_1) & R(x_2,x_2) & \dots & R(x_2,x_n) \\ \dots & \dots & \dots & \dots \\ R(x_n,x_1) & R(x_n,x_2) & \dots & R(x_n,x_n) \end{vmatrix} \tag{3.5}$$

among which $R(x_i,x_j) = |r_{ij}|$

From Equation 3.5 we can know that, $R(x_i,x_i) = 1$, $R(x_i,x_j) = R(x_j,x_i)$,

So this matrix satisfies Reflexivity and Symmetry.

Matlab code:

Algorithm 3.2 Construct fuzzy similarity matrix

```
e=corrcoef(d);
e=abs(e);
```

3.6.3 Construct fuzzy equivalent matrix and cluster

The fuzzy relationship matrix between the object and the object constructed by the above steps is generally just a fuzzy similarity matrix, which is not necessarily transitive, and thus may not be a fuzzy equivalent matrix.

Therefore, a fuzzy equivalent matrix must be constructed from the above fuzzy similarity matrix. The transitive closure of the fuzzy similarity matrix is a fuzzy equivalent matrix, and the data can be clustered based on the transitive closure. The specific steps are as follows:

Start from fuzzy similarity matrix R,we calculate the square one by one,see Equation 3.6

$$R \to R^2 \to R^4 \cdots \to R^{2^i} \to \cdots \tag{3.6}$$

When the equation $R^k \circ R^k = R^k$ exist,the R^k is the the fuzzy equivalent matrix.

After the fuzzy equivalence matrix is constructed, appropriately select the confidence level value λ. In this program, the λ value is the deduplication value of the fuzzy equivalent matrix. Matrix elements of the equivalence matrix greater than or equal to the λ value are classified into one cluster, when λ=1, no elements would be clustered, when $\lambda = min(equivalentmatrix)$, all elements would be clustered as one. A detailed explanation with example is in subsection 3.7.4.

Matlab code:

Algorithm 3.3 Construct fuzzy equivalent matrix

```
for i do=1:n;
    for j do=1:n;
        for k do=1:n;
            f(k)=min(e(i,k),e(k,j));
        end for
        g(i,j)=max(f);
    end for
end for
while (sum(sum(e == g))/(n^2)) = 1; do
    e=g;
    for i do=1:n;
        for j do=1:n;
            for k do=1:n;
                f(k)=min(e(i,k),e(k,j));
            end for
            g(i,j)=max(f);
        end for
    end for
end while
```

3.6.4 Feature screening

Through subsection 3.6.3 and the choice of λ, the data would be clustered into 3 classifications. For a classification that contains more than one financial indicator, a representative indicator can be selected by the correlation coefficient method. In the same category, the indicators with the greatest correlation with other indicators should be selected to ensure that the selected features can cover the most comprehensive information.

The detailed steps of feature screening:

1. Calculate the correlation coefficient between the indicators in each classification.
2. Calculate the correlation index, and select the indicator with largest correlation index. If there is only one indicator in the classification, it can be directly included in the final indicator set. If there are two indicators in the classification, we can randomly choose one (since the correlation index is equal) .

3.7 Result

3.7.1 The fuzzy similarity matrix

The financial profitability ratios, development capability, shareholders' equity, solvency, and operating capability of the listed company (The ratios calculation methodology are explained in Section 3.3 Financial Feature Calculation) are used to construct the fuzzy similarity matrix according to subsection 3.6.3. To not interrupt the logical continuity of the presentation, keep it concise and clear, we will only demonstrate the result of probability ratios of the year 2014. The results of the other 4 categories of ratios of 2014 are in Appendix A, which are based on the same procedure. The fuzzy similarity matrix of profitability of year 2014 is in Table 3.2.

3.7.2 The fuzzy equivalence matrix

The fuzzy equivalent matrix obtained by the transitive closure method according to subsection 3.6.3 are shown in Table 3.3.

3.7.3 Clustering

By assigning value of λ, the clustering results of profitability ratios are shown in Table 3.4.

The Fuzzy Similarity Matrix of Profitability Ratios									
	F1-1	F1-2	F1-3	F1-4	F1-5	F1-6	F1-7	F1-8	F1-9
F1-1	1	0.1299	0.0338	0.0377	0.0236	0.016	0.9134	0.0043	0.0474
F1-2	0.1299	1	0.8158	0.8179	0.8196	0.2042	0.3597	0.0261	0.0326
F1-3	0.0338	0.8158	1	0.9992	0.9904	0.1565	0.1586	0.0622	0.0355
F1-4	0.0377	0.8179	0.9992	1	0.9893	0.1587	0.1586	0.0635	0.0357
F1-5	0.0236	0.8196	0.9904	0.9893	1	0.1452	0.1618	0.0571	0.0149
F1-6	0.016	0.2042	0.1565	0.1587	0.1452	1	0.0278	0.0153	0.0109
F1-7	0.9134	0.3597	0.1586	0.1586	0.1618	0.0278	1	0.006	0.0244
F1-8	0.0043	0.0261	0.0622	0.0635	0.0571	0.0153	0.006	1	0.0121
F1-9	0.0474	0.0326	0.0355	0.0357	0.0149	0.0109	0.0244	0.0121	1

F1-1 Operating gross profit margin
F1-2 Operating net profit margin
F1-3 Return on assets before interest and tax
F1-4 Return on assets
F1-5 Return on current assets
F1-6 Return on fixed assets
F1-7 Marginal profit ratio
F1-8 Return on equity
F1-9 Growth rate of main business income

Table 3.2 The fuzzy similarity matrix of profitability ratios

The Fuzzy Equivalent Matrix of Profitability Ratios									
	F1-1	F1-2	F1-3	F1-4	F1-5	F1-6	F1-7	F1-8	F1-9
F1-1	1	0.3597	0.3597	0.3597	0.3597	0.2042	0.9134	0.0635	0.0474
F1-2	0.3597	1	0.8196	0.8196	0.8196	0.2042	0.3597	0.06357	0.0474
F1-3	0.3597	0.8196	1	0.9992	0.9904	0.2042	0.3597	0.06357	0.0474
F1-4	0.3597	0.8196	0.9992	1	0.9904	0.2042	0.3597	0.06357	0.0474
F1-5	0.3597	0.8196	0.9904	0.9904	1	0.2042	0.3597	0.06357	0.0474
F1-6	0.2042	0.2042	0.2042	0.2042	0.2042	1	0.2042	0.06357	0.0474
F1-7	0.9134	0.3597	0.3597	0.3597	0.3597	0.2042	1	0.06357	0.0474
F1-8	0.0635	0.0635	0.0635	0.0635	0.0635	0.0635	0.0635	1	0.0474
F1-9	0.0474	0.0474	0.0474	0.0474	0.0474	0.0474	0.0474	0.04749	1

F1-1 Operating gross profit margin
F1-2 Operating net profit margin
F1-3 Return on assets before interest and tax
F1-4 Return on assets
F1-5 Return on current assets
F1-6 Return on fixed assets
F1-7 Marginal profit ratio
F1-8 Return on equity
F1-9 Growth rate of main business income

Table 3.3 The fuzzy equivalent matrix of profitability ratios

The Clustering result of Profitability Ratios		
λ	Number of Classification	Class
1	9	{F1-1}{F1-2}{F1-3}{F1-4}{F1-5}{F1-6}{F1-7}{F1-8}{F1-9}
0.999235	8	{F1-1}{F1-2}{F1-3 F1-4}{F1-5}{F1-6}{F1-7}{F1-8}{F1-9}
0.990468	7	{F1-1}{F1-2}{F1-3 F1-4 F1-5}{F1-6}{F1-7}{F1-8}{F1-9}
0.913467	6	{F1-1 F1-7}{F1-2}{F1-3 F1-4 F1-5}{F1-6}{F1-8}{F1-9}
0.819606	5	{F1-1 F1-7}{F1-2 F1-3 F1-4 F1-5}{F1-6}{F1-8}{F1-9}
0.359789	4	{F1-1 F1-2 F1-3 F1-4 F1-5 F1-7}{F1-6}{F1-8}{F1-9}
0.204279	3	{F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7}{F1-8}{F1-9}
0.06357	2	{F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7 F1-8}{F1-9}
0.04749	1	{F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7 F1-8 F1-9}

Table 3.4 The clustering result of profitability ratios

3.7.4 Screen the ratios in same classification

As we explained in subsection 3.6.3, when $\lambda = 0.999235$, the elements F1-3 and F1-4 in the equivalent matrix equal to λ, so they can be clustered into the same cluster. Cluster based on this principle to the last value of λ. when $\lambda = 0.204729$, it forms 3 clusters.

We require 3 ratios from each financial aspects hence choose the classification when λ=0.204279 in Table 3.4, so the classification is F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7,F1-8 and F1-9, F1-8 and F1-9 are already single ratios in single classes so both can be count as one of the final 15 ratios. However, F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7 is in the same cluster, we need to choose one ratio that can represent the cluster by using correlation coefficient method.

1. Calculate correlation coefficient A_{ij}

Correlation Coefficients							
A_ij	F1-1	F1-2	F1-3	F1-4	F1-5	F1-6	F1-7
F1-1	1	-0.13	0.0338	0.0377	0.0236	0.0161	-0.9135
F1-2	-0.13	1	0.8159	0.818	0.8196	0.2043	0.3598
F1-3	0.0338	0.8159	1	0.9992	0.9905	0.1566	0.1586
F1-4	0.0377	0.818	0.9992	1	0.9894	0.1587	0.1586
F1-5	0.0236	0.8196	0.9905	0.9894	1	0.1453	0.1619
F1-6	0.0161	0.2043	0.1566	0.1587	0.1453	1	0.0278
F1-7	-0.9135	0.3598	0.1586	0.1586	0.1619	0.0278	1
F1-1 Operating gross profit margin							
F1-2 Operating net profit margin							
F1-3 Return on assets before interest and tax							
F1-4 Return on assets							
F1-5 Return on current assets							
F1-6 Return on fixed assets							
F1-7 Marginal profit ratio							

Table 3.5 Correlation coefficients

2. Calculate correlation index R

$$R_{A1} = a_{1,2}^2 + a_{1,3}^2 + a_{1,4}^2 + a_{1,5}^2 + a_{1,6}^2 + a_{1,7}^2 = 0.1424$$
$$R_{A2} = a_{2,1}^2 + a_{2,3}^2 + a_{2,4}^2 + a_{2,5}^2 + a_{2,6}^2 + a_{2,7}^2 = 0.3658$$
$$R_{A3} = a_{3,1}^2 + a_{3,2}^2 + a_{3,4}^2 + a_{3,5}^2 + a_{3,6}^2 + a_{3,7}^2 = 0.4493$$
$$R_{A4} = a_{4,1}^2 + a_{4,2}^2 + a_{4,3}^2 + a_{4,5}^2 + a_{4,6}^2 + a_{4,7}^2 = 0.4497$$
$$R_{A5} = a_{5,1}^2 + a_{5,2}^2 + a_{5,3}^2 + a_{5,4}^2 + a_{5,6}^2 + a_{5,7}^2 = 0.4466$$

$$R_{A6} = a_{6,1}^2 + a_{6,2}^2 + a_{6,3}^2 + a_{6,4}^2 + a_{6,5}^2 + a_{6,7}^2 = 0.0189$$
$$R_{A7} = a_{7,1}^2 + a_{7,2}^2 + a_{7,3}^2 + a_{7,4}^2 + a_{7,5}^2 + a_{7,6}^2 = 0.1735$$

3. Screening

   From the result we can conclude that R_{A4} is the largest correlation index, so F1-4 is picked into the final feature set.

   After screening, the 3 ratios in profitability category are: F1-8 Return on equity, F1-9 Growth rate of main business income, F1-4 Return on assets.

3.8 Summary

Profitability ratios	F1-4 Return on assets
	F1-8 Return on equity
	F1-9 Growth rate on main business income
Development capability ratios	F2-1 Appreciation rate of capital preservation
	F2-3 Fixed assets growth rate
	F2-5 Net profit growth rate
Shareholders' equity ratios	F3-5 Undistributed profit per share
	F3-6 Market to book ratio
	F3-9 EPS/P
Solvency ratios	F4-1 Current ratio
	F4-3 Debt ratio
	F4-7 Equity to total assets ratio
Operating capability ratios	F5-2 Inventory turnover
	F5-4 Working capital turnover
	F5-8 Total assets turnover

Table 3.6 Financial feature set

Table 3.6 is the clustering result of all 5 financial aspects of the year 2014. In conclusion, The fuzzy clustering analysis is used in this chapter to successfully reduced 42 financial ratios to 15 ratios in 5 categories. Each category contains 3 ratios. The 15 financial ratios and 9 technical indicators will be the input feature space of RFs and SVMs.

We are glad to see that some indicators that financial experts judged to be useful for stock price prediction based on prior experience were included in the final indicator set, such as ROE, PB, Net profit growth rate, Debt ratio, and Current ratio.

However, it is necessary to realize that financial experts believe these indicators are more effective based on long-term practical experience. They quantitatively or qualitatively estab-

lish a relationship between dependent variables (Stock Return) and independent variables (financial ratios).

In the clustering of FCA, there is no introduction of dependent variables, and the basis is only the internal mathematical relationship between variables. Of course, due to the prior validity, the calculation of the indicator set itself may tend to be calculated around more effective indicators, which influences our selection of included indicator sets. Apart from this passive intervention of prior validity, we will not actively intervene in the algorithm's behaviour (such as attribute reduction) based on prior experience in this study.

Chapter 4

Support Vector Regression

This chapter will introduce the basic theory of SVMs, and then will introduce three parameters that we will optimise. This chapter aims to enable readers to understand SVMs and the theoretic impact of the three optimised parameters on the regression results.

4.1 Rationale of Using Machine Learning

In machine learning, prediction problems could be divided into linear prediction and non-linear prediction problems. The prediction task of this research must be a non-linear prediction for the following reasons:

1. Paper have proved that the stock forecasting problem is a nonlinear problem, so the linear model is invalid [1].

2. If the stock forecast is a linear problem, then the forecast will be straightforward, contradicting reality.

Therefore, this thesis applies machine learning models such as support vector machine and random forest as nonlinear prediction models, excluding linear prediction methods such as multiple regression.

4.2 Rationale of Using RFs and SVMs

We choose RFs and SVMs as the prediction models for reasons:

1. Effectiveness

In the paper 'Do we need hundreds of classifiers to solve real world classification problems?' [14], the authors evaluate 179 classifiers arising from 17 families on 121 data sets from the whole UCI database and other real data sets.

The result shows that algorithms most likely to be the best are the random forest (RFs), the best of which achieves 94.1% of the maximum accuracy overcoming 90% in the 84.3% of the data sets. The advantage is not statistically significant with the second-best, the SVMs with Gaussian kernel, which achieves 92.3% of the maximum accuracy [14].

2. Background

When choosing a classification/regression model, we are limited by our conditioned background within computer science and mathematics [14].

The knowledge required by each machine learning method family is relatively fragmented. The theoretical foundation behind them come from different disciplines. for example, Neural Networks are closely related to linear algebra, while Genetic Algorithms imitate the natural selection process of genes. As purely statistical models, the naive Bayes model comes from the Bayesian principle, and the support vector machine comes from statistical learning theory.

Another fact supporting this argument is that great machine learning pioneers usually research one model for a long time but do not contribute much to other models, such as Leo Breiman, the inventor of random forests, and Geoffrey Hinton, the representative of neural networks.

I have some understanding and application experience of support vector machines and random forest models. It is another reason why I adopt these two models.

4.3 Non-technical Overview of SVMs

For a classification problem, firstly, we consider the case of two-dimensional data (that is, the position of the predicted value is determined by two variables). The red line in Figure 4.1 (the middle line with 45 degree slope in white/black printing) is the optimal separation line.

In three-dimensional data (where the predicted value is determined by three variables), two classes of data can be separated by an optimal separation plane.

In multi-dimensional data, data can be separated by an $n-1$ (n is the number of dimensions) dimensional plane, which is called the separation hyperplane.

How is the position and direction of this line determined?

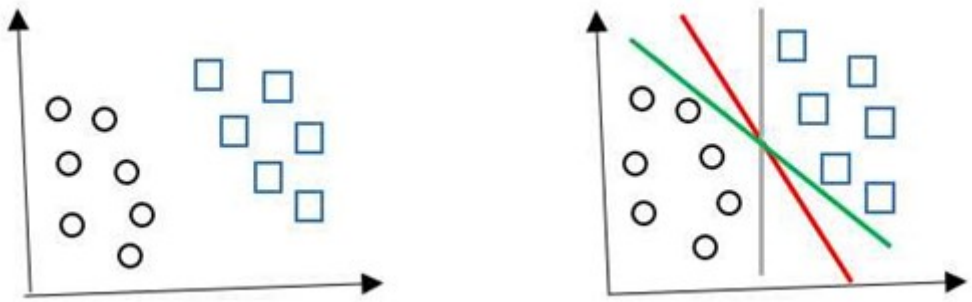


Fig. 4.1 Hyperplanes [17]

The principle is very similar to OLS least squares regression. As shown in Figure 4.1 in Section 4.5, draw two dotted lines along with the closest sample points on both sides of the hyperplane, and the vertical distance between them is called the minimum interval. The hyperplane with the maximum interval is the optimal solution SVMs is looking for. The sample points crossed by the dotted lines on both sides corresponding to this optimal solution are the support sample points in the SVMs, which are called 'support vectors'.

When the geometric properties of the hyperplane are fixed as 'flat' (two-dimensional: straight line, three-dimensional: plane), the performance of SVMs is limited. Through the kernel function, we can map the plane into a curved surface, thereby significantly improving the performance of SVMs.

In our program, we will optimise 3 parameters of the SVMs: C, σ and ε.

C is the penalty term. When C takes a finite value, SVMs allows some samples drop in the wrong side of the separation of the hyperplane.

σ is a parameter in the radial basis kernel, it is the kernel width that affects the complexity of the sample data in the high-dimensional feature space.

ε is the slack term. When we apply SVMs to regression problem (Support Vector Regression, SVR), SVR assume that we can tolerate a maximum bias ε, that is, we only count the sample individual in loss when the absolute value of the difference between $f(x)$ and y is larger than ε.

4.4 Support Vector and Margin

Section 4.4 to 4.7 are detailed explanations of basic technical details and processes in SVMs.

Given a training sample set, the primary task of classification is to find a hyperplane in the sample space based on the training set and separate samples into different classes.

There may be many hyperplanes that can separate training samples. Which one should we choose? Intuitively, we should find the hyperplane in the 'middle' of the two types of training samples, that is, the red one in Figure 4.1.

This should be the hyperplane that has the best tolerance for disturbances of the training samples, for example, due to the limitation of training set or data noise, the samples outside the training set (i.e. testing set) may be closer to the separation boundary between the two classes than the training samples in Figure 4.1. This will make many hyperplanes (when separate the testing set) make mistakes, and the red hyperplane will be the least affected. In other words, this hyperplane is the most robust and has the strongest generalization ability [22].

In the sample space, the hyperplane can be described by the following linear equation:

$$w^T x + b = 0 \tag{4.1}$$

Where $w = (w_1; w_2; ...; w_d)$ is the normal vector, decides the direction of the hyperplane, b is the displacement term, decides the distance between the hyperplane and the origin. Apparently, the hyperplane could be decided by w and b, denote it as (w, b), the distance between a random point x to the hyperplane (w, b) is:

$$r = \frac{|w^T x + b|}{\|w\|} \tag{4.2}$$

Assume that the hyperplane could classify the training sample correctly ,that is when $(x_i, y_i) \in D, if y_i = +1, w^T x_i + b >= 0. if y_i = -1, w^T x_i + b <= 0$ then:

$$\begin{cases} w^T x_i + b >= +1, y_i = +1 \\ w^T x_i + b <= -1, y_i = -1 \end{cases} \tag{4.3}$$

As shown in Figure 4.2, the training sample points closest to the hyperplane make the equal sign of Equation 4.3 hold. They are called support vectors. The sum of the distances from the two heterogeneous support vectors to the hyperplane is:

$$\gamma = \frac{2}{\|w\|} \tag{4.4}$$

which is called margin.

4.5 Soft Margin

In the previous discussion, we assumed that the training samples are linearly separable in the sample space or feature space, that is, there is a hyperplane that can completely separate different classes of samples. However, it is usually difficult to find a appropriate hyperplane that makes the training sample linearly separable in the real tasks. Even if a certain hyperplane is found to make the training set linearly separable in the feature space, it is difficult to conclude that this seemingly linearly separable result is not caused by over-fitting.

One way to alleviate this problem is to allow the support vector machine to make errors on some samples. For this reason, the concept of soft margin should be introduced, as shown in Figure 4.2.

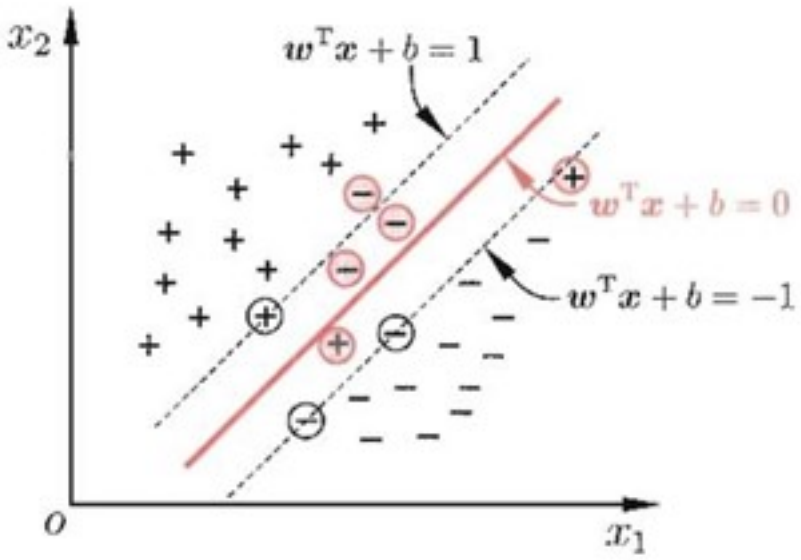


Fig. 4.2 Soft Margin [17]

Specifically, the previously introduced support vector machine requires all samples to meet constraint Equation 4.3, that is, all samples must be divided correctly, which is called a 'hard margin', while a soft margin allows some samples to not meet the constraint:

$$y_i(w^T x_i + b) \geq 1 \tag{4.5}$$

Apparently, while maximizing the interval, the samples that do not satisfy the constraints should be as few as possible, so the optimisation goal can be written as

$$[min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} l_{0/1}(y_i(w^T x_i + b) - 1)] \tag{4.6}$$

where $C > 0$ is a constant, $l_{0/1}$ is '0/1 loss function'

$$l_{0/1}(z) = \begin{cases} 1, if\ z < 0; \\ 0, otherwise. \end{cases} \tag{4.7}$$

Obviously, when C is infinite, Equation 4.6 forces all samples to satisfy the constraint 4.5, then Equation 4.6 is equivalent to the original form without tolerance to the sample not satisfy the constraint. When C takes a finite value, Equation 4.6 allows some samples not to satisfy the constraint, which give the SVMs a soft margin.

C is the first parameter we will optimise. Referring to the relevant paper's recommendation, we set the search range C to [0.1, 10] [23].

4.6 Kernel Function

Another technique to solve the problem of sample linear inseparability is kernel function mapping. The sample can be mapped from the original space to a higher-dimensional feature space, so that the sample is linearly separable in the feature space. If the original space is finite-dimensional, that is, the number of features is limited, then there must be a high-dimensional feature space which makes the sample separable. This dimension raising process is carried out through the kernel function. In this research, we use the radial basis kernel function. The formula is:

$$K(x, x_i) = \exp\{-\frac{|x - x_i|^2}{\sigma^2}\} \tag{4.8}$$

Obviously, the radial basis kernel function introduces the parameterσ. The kernel widthσ affects the complexity of the sample data in the high-dimensional feature space [24].

σ is the second parameter we will optimise. It can be seen from the formula that the size of σ^2 is relative to the value of $|x - x_i|$. Therefore, in practical application, as long as the value of σ^2 is much smaller than the minimum distance $\min|x - x_i|$ between the training samples, the effect of small σ^2 can be achieved. The effect of $\sigma^2 \to \infty$ can be achieved while σ^2 is much larger than the maximum distance between training samples. Based on this consideration, the optimal search scope σ^2 is:

$$[\min(|x - x_i|)^2 \times 10^{-3}, \max(|x - x_i|)^2 \times 10^{-3}] \tag{4.9}$$

Studies have shown that serious 'over-learning' phenomenon occurs when $\sigma^2 \to 0$, and the model does not have any generalization ability on the test sample, in contrast when

$\sigma^2 \to \infty$, serious 'under-learning' phenomenon will occur. So the search space of σ^2 should be pre-estimated [22].

In the empirical part of this thesis, considering the calculation complexity of above optimal search range, we set the search range to [0.1,10].

4.7 Support Vector Regression

Now let's consider the regression problem, given a training sample

$D = (x_1, y_1), (x_2, y_2), ..., (x_m, y_m), y_i \in R$,

We hope to derive a regression model

$$f(x) = w^T x + b \tag{4.10}$$

so that $f(x)$ and y are as close as possible. w and b are the model parameters to be determined.

For sample (x, y), the traditional regression model usually calculate loss based on the difference between model output $f(x)$ and real value y, if and only if $f(x)$ is equal to y, the loss is 0. Different from this, SVR assume that we can tolerate a maximum bias ε, that is, we only count it in loss when the absolute value of the difference between $f(x)$ and y is larger than ε. As shown in the Figure 11 below, this actually constructs a margin with width 2ε, if the training sample fall into the margin, we still count it as the correct prediction [25], hence ε is the third parameters we will optimise. With reference recommended and the complexity of program implementation, it is determined that the search range of ε is [0.01.1] [23].

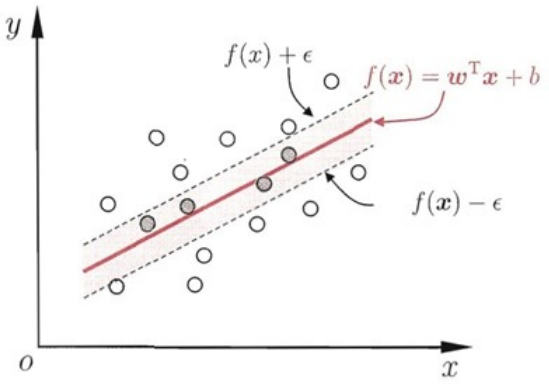


Fig. 4.3 Support vector regression

Hence, the SVR problem can be written as

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i-1}^{m} l_\varepsilon(f(x_i) - y_i) \tag{4.11}$$

where C is the regularization constant, l_ε is the ε-insensitive loss function:

$$l_\varepsilon = \begin{cases} 0, & if|z| \geq \varepsilon \\ |z| - \varepsilon, otherwise. \end{cases} \tag{4.12}$$

4.8 Summary

This chapter explains one of the main models of this study: the support vector machine. First we explained the motivation for applying the machine learning model, followed by a non-technical explanation of SVMs and three important parameters for the reader's reference. Then, beginning with two-dimensional classification, we explained the dimensionality raising process of SVMs, which naturally introduces three important variables. The classification and regression are different applications of support vector machines but they share the same principles. The next chapter mainly introduces another machine learning method - Genetic algorithm to optimise the three important variables mentioned in this chapter.

Chapter 5

Parameter Optimisation

According to the pipeline in Figure 1.4, in this chapter, we will introduce the parameter optimisation of the GA-SVR. In Chapter 4, we introduced the details of support vector machines and the optimisation range of three parameters. We will use genetic algorithm (GA) to optimise the parameters of support vector machines.Hence In this chapter, we will demonstrate the technical details of GA.

In essence, this chapter describes how to use a genetic algorithm to optimise the 3 parameters (C,σ,ε) to maximise the performance of the SVR.

5.1 Reason of Choosing GA for Parameter Optimisation

5.1.1 Combination explosion

Combination explosion means that with the increase of the optimisation parameter dimension, the different combinations of parameter variable values increase exponentially.

In this thesis, we have to optimise three parameters. Assume each parameter has 100 possible values then we have 100^3 possible combinations. Although it is not challenging to perform a grid search on 100^3 combinations with available computing power, we decided to apply a parameter optimisation method considering computational efficiency and potential larger parameter scale.

5.1.2 Effectiveness

Plenty of literature has applied the genetic algorithm to optimise SVMs and proved that GA is an efficient method [26, 27]. Some literature proved that GA-SVRs is an efficient combination to complete the regression objective [28, 29]. The above literature has been

reviewed in Chapter 2. The effectiveness of GA optimisation in this research is verified in Subsection 7.8.3.

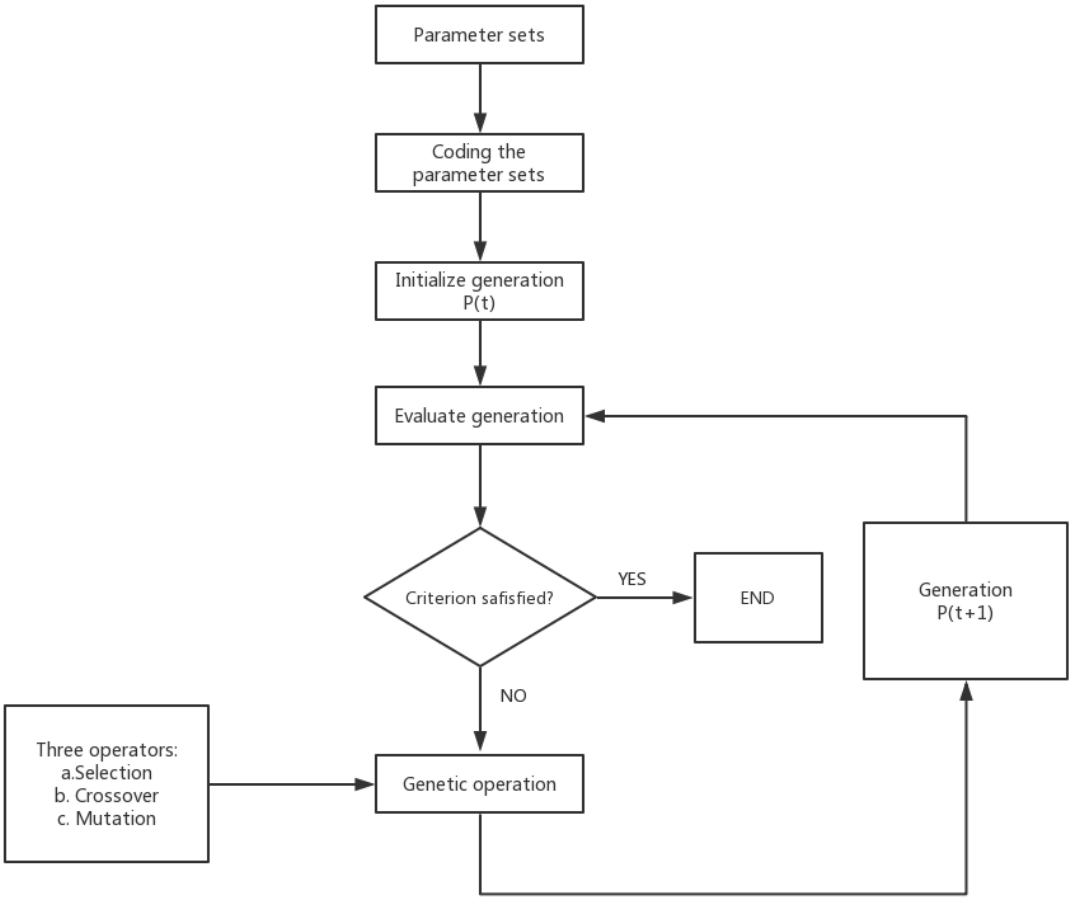


Fig. 5.1 The basic flow of Genetic Algorithm

The basic flow of a simple genetic algorithm can be summarized as shown in Figure 5.1, that the genetic algorithm is an iterative loop to implement the search process. The steps that need to be completed are [30]:

1. Choose the coding strategy and design fitness function $f(x)$. Determine specific genetic strategies, including population size n, selection, crossover and mutation operators p_c and their trigger probability p_m.
2. Complete the encoding operation, randomly generate the initialization group P.
3. Calculate the fitness $f(x)$ for each individual in the population.

4. Use selection, crossover, and mutation operators for genetic manipulation to form the next generation of populations.

5. Judge whether the fitness value satisfies the predetermined criteria. If not, return to the previous step.

5.2 Parameter Encoding

Mapping the problem to be solved to the coding space is called encoding. There are three kinds of coding schemes: binary coding, real number coding, and floating-point coding. The selection of the coding scheme generally depends on the nature of the problem to be solved. In this research, we use floating-point encoding.

5.3 Fitness Function

After the genetic algorithm mapping the problem to the coding space, in order to implement the principle of survival of the fittest, the environmental adaptability of each solution (chromosome) must be evaluated.

Fitness function is the indicator to evaluate the chromosome and is the objective function of the optimisation problem, which can guide the direction of population evolution. Generally, Mean Squared Error (MSE) is a proper fitness function for a regression problem.

5.4 MSE

MSE is short for mean squared error and is the average distance between the predicted and true values. The formula is as follows:

$$MSE = \frac{\sum_{i=1}^{1} (a_i - \hat{a}_i)^2}{N} \tag{5.1}$$

N represents the number of training set samples, a_i is the predicted value, $\hat{a}_i$ is the true value. Large MSE indicate low accuracy of the prediction, vice versa.

5.5 Genetic Operators

Genetic algorithms use three types of genetic operators to simulate population evolution, shown in Figure 5.2. The selection operator is used to simulate the survival of the fittest

mechanism in nature. The crossover operator is used to simulate the breeding mechanism. The mutation operator is used to simulate the mutation phenomenon.

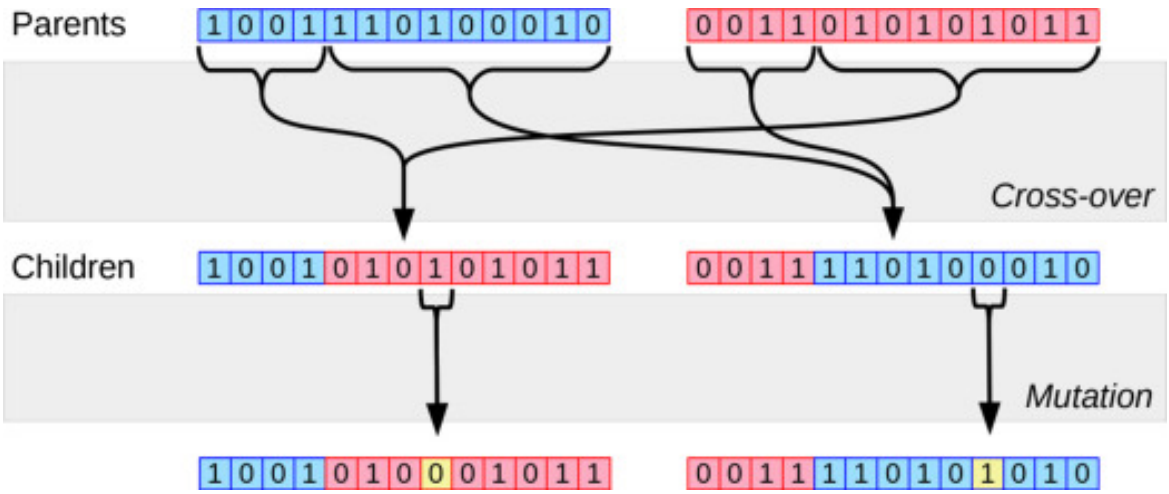


Fig. 5.2 Selection, Crossover and Mutation

- Selection

  The purpose of selection is to select competent individuals from the population, allowing them to reproduce their offspring as parent generation. The key is to sort individuals based on their fitness values and select the superior ones. They will be inherited from population $P(t)$ into the next generation population $P(t+1)$.

- Crossover

  Crossover is to randomly match each individual in the population $P(t)$ into pairs, and then exchange part of chromosomes between individuals according to a certain crossover probability.

- Mutation

  The mutation operation is to replace the value of some locus in the individual chromosome string with other value of the locus. The main reason for using mutation operators in genetic operation is that it can improve the local search ability of genetic algorithms and maintain the diversity of the population.

5.6 Construction of GA-SVR

When applying the genetic algorithm to SVR parameters optimisation, the basic steps of the algorithm are shown in Figure 5.3:

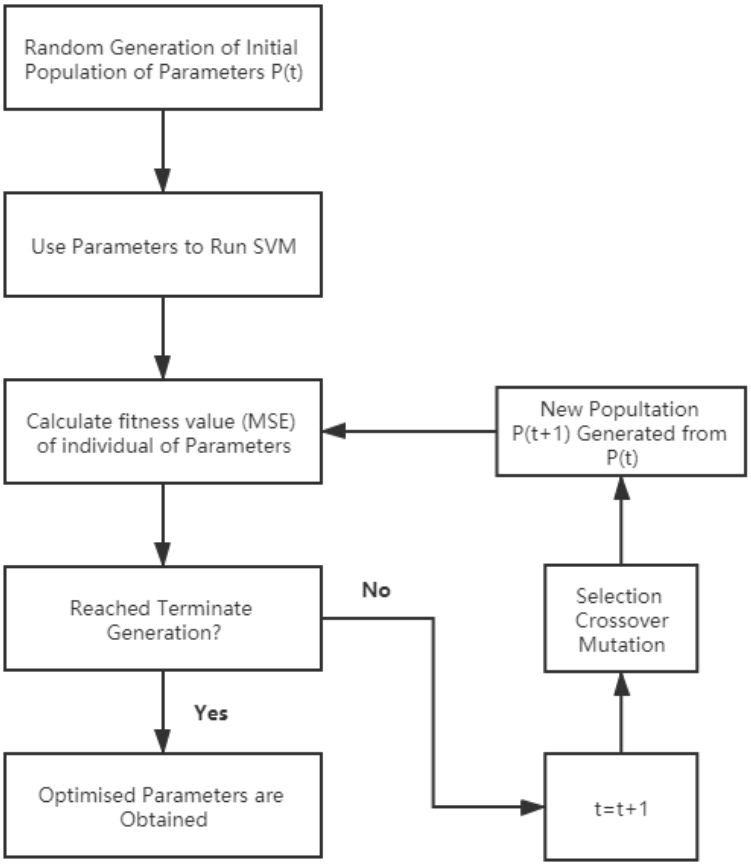


Fig. 5.3 The Pipeline of Genetic Algorithm - Support Vector Machine

Figure 5.3 shows that optimising SVR with GA essentially takes the evaluation function of SVR as the evaluation function of GA. It is the critical point to combine the two algorithms. SVR is called cyclically in the GA framework and the optimal parameters are selected from the last generation of parameters. The Terminating Generation of GA-SVR is a fixed value T, the algorithm stops when $t = T$. The value of T in this research is explained in section 5.7.

5.7 GA Hyper-Parameters Setting

In the genetic algorithm, deciding control parameters (hyper-parameters) is crucial. Hyper-parameters will affect the performance of the genetic algorithm. Therefore, appropriate parameter setting should be implemented based on the prior experience.

The hyper-parameters involved in the whole genetic operation include population size N, terminated evolutionary generation T,crossover probability P_c, and mutation probability P_m . Scholars have carried out relevant research and given practical suggestions on the selection of optimal hyper-parameters [26].

- Population size N

  Population size N: population size will directly affect the genetic algorithm's convergence speed and search efficiency. The choice of population size is to find a balance between algorithm efficiency and algorithm effectiveness, which can be obtained simply by observing the changes in the evaluation result of both aspects as the population size increases. Referring to related literature and the observation of the experiment, the population size in practical application is generally taken between 20 and 200. in this thesis, N is set to 100.

- Terminated evolutionary generation T

  The principle of determining terminated evolutionary generation T is similar to population size. T is generally set as 10 to 500. In this research, T is 20.

- Crossover probability P_c

  The crossover probability P_c controls the probability of the crossover operator being triggered, it will seriously affect the final performance of the algorithm. The larger the P_c, the more thoroughly crossover will be performed. However, if the value of P_c is too large, then the frequent updating genes will increase the probability that missing gene with high enough fitness, while a small P_c will allow more genes to be copied directly to the next generation, which will lead to search block. P_c is usually between 0.4 to 0.9, in my thesis it is 0.7.

- Mutation probability P_m

  The mutation probability controls the frequency that the mutation operator is triggered at the end of the crossover operation, each gene will be mutated randomly according to the mutation probability P_m. Although the mutation probability can increase the diversity of the population, if it is too large, the genetic algorithm will approximate a random search algorithm. P_m is generally between 0.001 and 0.01, in this research, $P_m = 0.7/lind$, where $lind$ is the chromosome length. It is the default setting of the Matlab GA function.

This chapter described the principle and parameter settings of Genetic Algorithm used to optimise the Support Vector Regression parameters, coupled with SVR in Chapter 4, formed GA-SVR. Next chapter will introduce another model of the thesis - Random Forest Regression.

Chapter 6

Random Forest Regression

In this chapter, we will introduce the principles of the random forest algorithm. Section 6.1 and 6.2 will introduce decision trees, Section 6.3, 6.4 and 6.5 will introduce how to use decision trees to form an ensemble random forest based on bagging principle.

6.1 Decision Tree

The forest is made up of 'trees' - decision trees, which is a simple type of machine learning.

Taking the binary classification task as an example, we hope to obtain a model from a given training data set to classify a testing data set. This task of classifying samples can be regarded as a decision process for the question of 'Is the current sample in the positive class?'.

Decision trees are based on a tree structure to make decisions. This is a natural processing mechanism for humans when facing decision-making problems. When making decisions on such issues, we usually make a series of judgments or 'sub-decisions'. For example, if we have to decide, 'Is this a delicious apple?', usually we will look at 'What color is it? ', if it is 'Red', then look at 'What is the state of its skin? ', if it is 'Fresh', we come to the final decision: this is a delicious apple, the decision process is shown in Figure 6.1:

Obviously, the conclusion of the decision-making process corresponds to a series of judgment results. Each judgment question raised in the decision-making process is a test of a particular feature. The result of each test leads to the conclusion or deriving further judgment questions, whose scope of consideration is within the limited scope of the last decision result.

A decision tree contains a root node, several internal nodes, and several leaf nodes as shown in Figure 6.1. The leaf nodes correspond to the decision results, and other nodes correspond to feature tests. The sample-set contained in each node is divided into sub-nodes based on the result of the feature test.

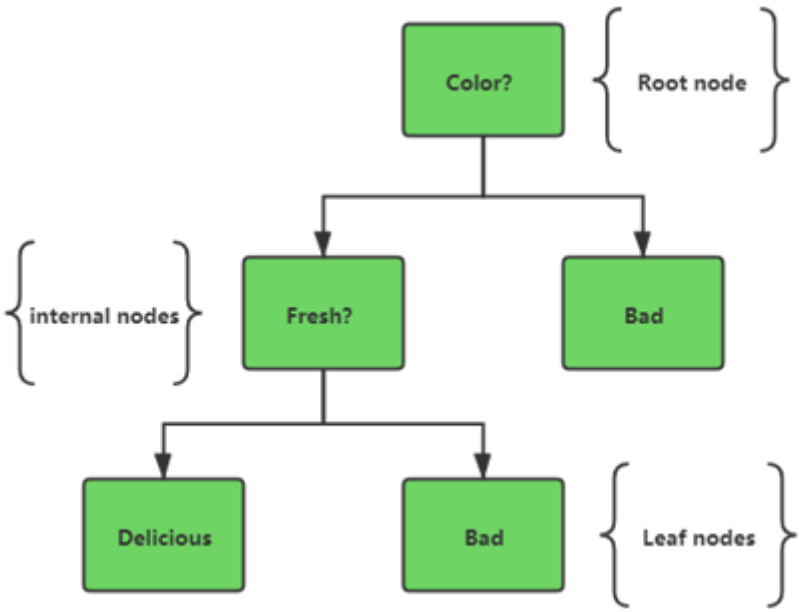


Fig. 6.1 Decision Tree Process

The root node contains the complete set of samples. The path from the root node to each leaf node corresponds to a decision test sequence. The objective of training a decision tree is to produce a decision tree with solid generalization ability, that is, the ability to deal with unseen examples. Its basic process follows a simple and intuitive divide and conquers strategy. The pseudo-code of the decision tree is as Algorithm 6.1:

The key to the algorithm is the 8th line: how to choose the optimal partition feature. As the partitioning process proceeds, we hope that the samples in the branch nodes belong to the same class as possible. That is, the purity of the node becomes higher and higher. The function to measure purity is called the purity function. The most commonly used purity function in classification tasks is information entropy.

6.2 The Process of Decision Tree Regression

In Section 6.1, we use a classification example to illustrate the operation of decision trees. There are only two primary differences between regression tree and classification tree.

1. Different indicators for measuring branch results: different purity functions, classification trees use information entropy and Gini impurity, regression trees use MSE, MAE, and other indicators suitable for regression.

Algorithm 6.1 The pseudo-code of the decision tree

Input: Training set $D = (x_1, y_1), (x_2, y_2), ..., (x_m, y_m)$; Feature set $A = a_1, a_2, ...a_d$
Process: Function TreeGenerate (D,A)

```
Generate node
if all samples in D belong to the same category C then
    Mark node as leaf node; return
end if
if A = Φ, OR the samples in D have the same value on A then
    Mark node as a leaf node, and mark its category as the category with the most samples in D; return
end if
Select optimal partition feature a_*, from A
for each value a_*^v do
    generate a branch for node, let D_v be sample subset with value a_x^v on a_* in D
    if D_v is empty then
        mark node as leaf node, and mark its category as the category with the most samples in D; return
    else
        TreeGenerate (D_v, A \ a_m)
    end if
end for
```

Output: A decision tree with node as root node

2. Regression result is essentially a numerical class. The leaf nodes of the classification tree are divided into classes. In contrast, at the leaf nodes of the regression tree, the average value will be calculated based on the numerical results in the divided classes, and this average value is the regression value of the leaf node.

6.3 Ensemble Learning

Ensemble learning accomplishes the learning task by constructing and combining multiple learners, sometimes called a multi-classifier system or committee-based learning. Ensemble learning by combining multiple learners can usually obtain significantly better generalization performance than a single learner, especially for weak learners. Therefore, many theoretical studies of ensemble learning are conducted on weak learners, and base learners are sometimes directly called weak learners.

The general structure of ensemble learning is to generate a set of 'individual learners' and then use some strategy to combine them. The individual learners are usually generated from an existing learning algorithm, for example, the C4.5 decision tree algorithm, BP neural network algorithm, etc. The ensemble algorithm usually contains the same type of individual learners, for example, the ensemble decision tree consists of decision trees, and the ensemble neural network consists of neural networks. Such an ensemble algorithm is 'homogeneous'. The individual learner in homogeneous ensemble learner is also called the 'base learner', and the corresponding learning algorithm is called the 'base learning algorithm'.

The ensemble learners can also include different types of individual learners, such as a decision tree and a neural network at the same time. Such an ensemble is 'heterogeneous'. The individual learners in a heterogeneous ensemble are generated by different learning algorithms, and there is no base learning algorithm. Correspondingly, individual learners are generally called 'component learners'.

According to the generation method of individual learners, the current ensemble learning methods can be roughly divided into two categories.

1. There is a strong dependency between individual learners, so they must run serially.

2. There is no strong dependency between individual learners so that they can run in parallel.

The representative of the latter is Bagging and Random forest.

6.4 Bagging

To obtain an ensemble learner with strong generalization performance, the individual learners in the ensemble should be as independent as possible. Although it cannot be independent of the actual data set, we can try to make the base learner as different as possible.

One possible approach is to sample the training data to generate several different subsets and then train base learner from a different subset of the training data. However, in order to obtain a better ensemble, we also hope that the individual learners should not perform ineffectively. If each subset of data is entirely different, then each base learner uses only a tiny part of the training data, which is not even effective learning. In order to solve this problem, we can consider using bootstrap sampling subsets.

Bagging is directly based on bootstrap sampling:

Given a data set containing N samples, we randomly take out a sample and put it into the sampling set, and then put the sample back into the initial data set, so that the sample may still be selected in the next sampling. In this way, after N times sampling operation, we get a sample set containing N samples. Some samples in the initial training set appear multiple times in the sample set, and some never appear.

In this way, we can sample T sample sets containing N training samples, then train a base learner based on each sample set, and then combine these base learners. This is the basic process of bagging. When the output is combined, bagging usually uses the voting method for classification tasks and the average method for regression tasks to get the final result.

6.5 Random Forest

Random Forest (RF) is an extended variant of bagging. On the basis of building bagging ensemble based on the decision tree, RFs further add random feature selection in the training process of the decision tree.

Specifically, the original decision tree selects an optimal feature from the feature set of the current node. In RFs, for each node of the base decision tree, a subset containing features is randomly selected from the feature set (contains features)of the node, and then an optimal feature is selected from this subset. Define the parameter k , which controls the degree of features selection randomness. In general, the recommended value $k = log_2 d$, where d is the number of features in the features set on the current node.

Random forest is simple and efficient. It exhibits powerful performance in many tasks. Random forest only makes small changes to 'bagging', but unlike bagging which the diversity of base learners is only through sample perturbation, the diversity of base learners in the

random forest comes not only from sample perturbation but also from feature perturbation. So that the generalization performance of the final ensemble can be further improved by increasing the degree of difference between individual learners.

Chapter 7

Result and Evaluation

In Chapter 3, we introduced the input data and data engineering of this research. In Chapters 4 and 5, we introduced the operating process and related optimization methodology of the GA-SVR. In Chapter 6, we introduced the Random Forest. We will analyse the output results after entering the data into the model in this chapter.

7.1 Program and Data Acquisition

7.1.1 Program

The research program is implemented in Matlab 2018a, where:

The Fuzzy clustering part is written based on the principle of the transitive closure method which was defined in Section 3.6.

Random forest is written and revised on the basis of the built-in Matlab functions 'regRF_train' and 'regRF_predict'.

Support vector machine part is written and revised on the basis of the built-in Matlab function 'libsvm'.

Genetic algorithm optimisation is written and revised on the basis of the built-in Matlab function 'ga'.

7.1.2 Data source

The financial ratio data is calculated from the annual report data of the listed company downloaded from the China Stock Market Accounting Research Database (CSMAR). The data required for stock prices, trading volume, and technical indicators are downloaded

from the 'Tongdaxin' stock trading platform. The access of two platforms and detailed data acquisition process can be found in Appendix B.

7.2 Data and Rolling Window Method

As we mentioned in Section 3.3, each year's financial data are published by the listed company no later than 30th April of the next year. Hence we will discretely match the financial data and stock returns to avoid using future data. For example, the result of the year 2014 is actually the result of testing on returns between 1st May 2015- 1st May 2016 by using the financial ratios of the year 2014 (collecting before 30th April 2015). This is a very important issue to avoid confusion of the result of the thesis, each time we see '2014', the corresponding financial indicators is from 2014 and the corresponding prediction period is '1st May 2015 - 1st May 2016'.

A rolling window method is applied, the length of the window is 10 days, that is, on day j, the training input X is the input features of day $j-9$, the training label Y is the stocks' 10 days return of day $j-9$, which is recorded on day j, the testing input X is the features of day j, and the predicting result is the stock's 10 days return of day j, which will be verified on day $j+9$. The rolling step length is 1 day.

For each day, the input of the program is a matrix of n rows and m columns, and the output is n rows of return, n is the number of input features calculated in Chapter 3, contains both technical indicators and financial indicators. m is number of the available stocks trading on that day.

7.3 Reasons of Result Transformation

For each day, the algorithm will generate n results, n is the number of stocks on the trading day. We need to transform this result to an evaluation-able form for two reasons:

1. n is an enormous, unconstant number. On each day there are thousands of stocks in trading and some stocks are suspended. When some major events occur, the number of suspended stocks may reach hundreds, and there will be hundreds of output results less than usual on this day. In order to continuously evaluate the effectiveness of the output of the algorithm on a daily basis, obviously we need to eliminate this gap.

2. Due to the volatility of the whole market, there will be a big difference between the overall predicted value and the overall actual value (which is consistent with the analysis in Subsection 2.5.4).

For example, if the market was in a bull market 10 days ago, but the market fell today, then today's predicted return based on the data 10 days ago will be higher overall. Since the purpose of this research is not to predict the overall return of the market, but to predict the return of individual stocks, in other words, our purpose is to predict which stocks are worth buying and which stocks are worth selling, so we just want to know the stocks relative return contrast with other stocks, hence it is necessary to eliminate this systematic difference due to the market volatility.

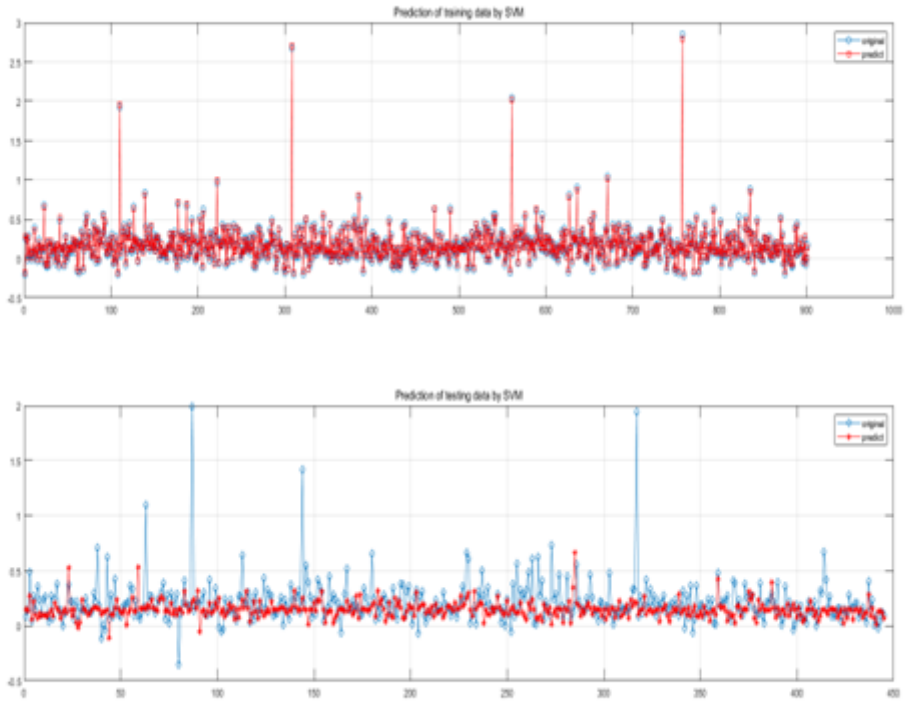


Fig. 7.1 An example of SVR training and predicting result for one day

Figure 7.1 is an example that illustrates the above two reasons, the X axis is the stocks number (stock 1, stock 2, etc.), Y axis is return. The upper half figure is the training result of the training stage of day 1, the lower half figure is the predicting result of the testing stage of day 11, the blue square dots are the actual return, the red dots are the predicting return generated by the algorithm. Other than an arbitrary stock numbering system, there is no logical ordering of the stocks along the X axis in Figure 7.1.

From the Figure 7.1, we can find that:

1. Respect to reason 1, the number of available stocks in trading in the upper figure is clearly lower than that of the lower figure, which shows that the number of stocks on trading of different days is not equal

2. Respect to reason 2, the average of the predicting returns is smaller than the actual return, which would be observed on day $j+10$, and more extreme values appear in the real data, which is hard to predict accurately.

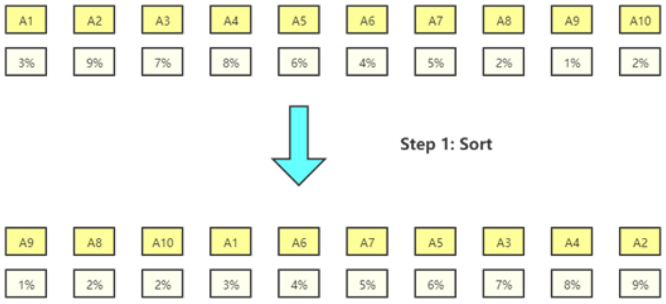
7.4 Result Transformation

In order to solve the problem in Section 7.3, we will take the following steps to transform the original result into a more comparable result.

Assume that we have 10 stocks: A1, A2, A3, A4, A5, A6, A7, A8, A9, A10

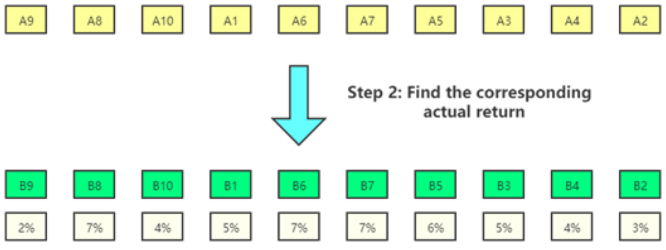
Step 1: Sort

Sort the stocks' predicting returns of each day in ascending order



Step 2 Find the corresponding actual return

In other words, we sort the actual returns based on the predicting returns



Step 3 Allocate actual return in groups, the number of groups is fixed

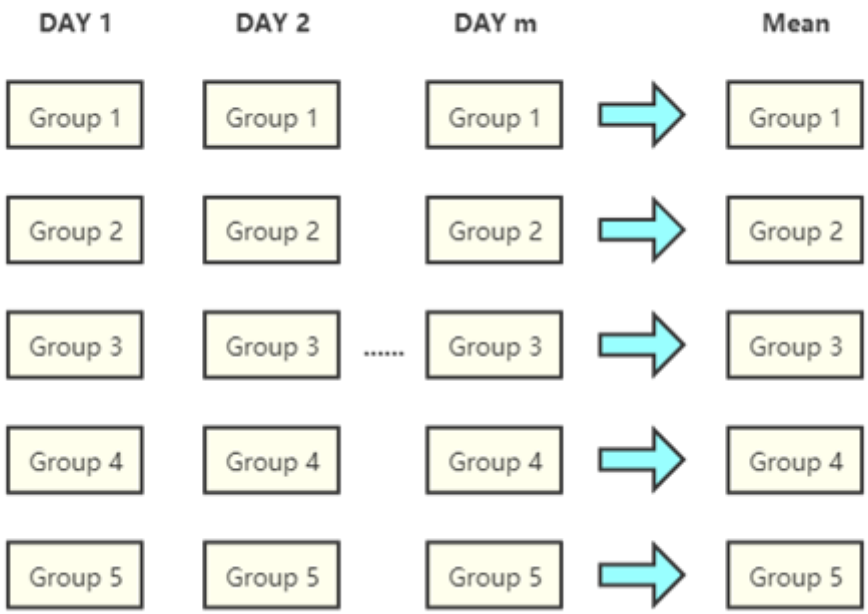
We chose a fixed group number and allocated the actual returns of each day into the groups. In our research, the group number is 50, while in this example, the number is 5.

Step 4 Calculate average return of each group

After dividing the returns into groups, the mean of the returns of each group is calculated. For example, the mean of Group 1 is (B9 + B8)/2 = 4.5%.

Step 5 Calculate average return of one year of each group

For each testing day in one year, repeat step 1 to 4, we will get a $l * m$ matrix, among which l is the number of groups, m is the number of testing days, since through step 2 the number group of each day is constant now, so we can calculate the mean of the same group of all days in that year.



Finally, through transformation, we get a column of average returns of all groups of each year. The Matlab code of Section 7.4 is shown in Algorithm 7.1

Algorithm 7.1 Data standardization

```
Allo=500; % How many stocks the capital would be invested in
H=zeros(50,size(a,2));
for i1 =1:size(a,2) do
    b=a(:,i1);
    c=b( isnan(b));
    j=ceil(length(c)/50);
    d=buffer(c,j);
    h=mean(d)';
    L1=d(:,end);
    if sum(L1) =0; then
        h(end)=mean(L1(L1 =0));
    else  h(end)=0;
    end if
    H(1:length(h),i1)=h;
end for
for u=1:size(H,1) do
    Z=any(H(u,:)==0);
    if Z==1 then
        H9=H(u,:);
        insertave=mean(H9(H9 =0));
        H9(H9==0)=insertave;
        H(u,:)=H9;
    end if
end for
```

7.5 Result of Two Models

7.5.1 Visualized result

Through transformation, we get a column of average returns of 50 groups for each year. According to step 1 and step 2 in Section 7.4, if our prediction is valid, that is, if we predict a group with a higher average return and obtain a higher average return, then the return of the 1 to 50 groups should be an ascending sequence.

If we use the $y-axis$ as the return of each group and the $x-axis$ as the group number to establish a coordinate system, the image should roughly have an upward trend if the prediction is valid. Visualisations of this kind of images are presented below in Figures 7.2, 7.3, 7.5, 7.6, 7.7, 8.1, 8.2.

7.5.2 Return prediction classification accuracy (RPCA)

The general evaluation indicator for regression should be Mean Squared Error (MSE). However, MSE is not suitable here due to the reasons we discussed in Section 7.3. Since we only care about the relative ranking of each stock (or group), if our prediction is effective, then the stocks we predicted have a higher return should have an actual higher return. In the most optimal circumstance that the prediction of all stocks' relative rankings are correct, the correlation coefficient between the predicted group ranking number and the real group ranking number should be 1, however, we can not get this optimal circumstance because of the statistical rejection caused by outliers and the small number of groups.

We loosen the standard and invent an measure called Return prediction classification accuracy (RPCA) to qualitatively measure the number of correct predictions, i.e. the number of true positives and true negatives (accurately classified as above and below mean return). It defined proportion of return predictions which are classified accurately as above or below average.

Define A_i,

$$A_i = \begin{cases} 0, (R_p - M_p) * (R_a - M_a) < 0 \\ 1, (R_p - M_p) * (R_a - M_a) \geq 0 \end{cases} \tag{7.1}$$

where, $i = 1, 2, 3...T$, i is the number of groups, T is the number of groups. A_i is a binary value, R_p is the predicted average return of the group i, M_p is the predicted average return of all the groups. R_a is the actual average return of the group i , M_a is the actual average return of all the groups.

Then:

$$RPCA = \frac{\sum_{i=1}^{T} A_i}{T} \tag{7.2}$$

RPCA can quantitatively measure the effectiveness of the algorithm on predicting the groups' rankings.

We can understand this measurement indicator from the perspective of confusion matrix in Table 7.1.

		Actual	
		Positive	Negative
Predicted	Positive	True positive	Type I Error
	Negative	Type II Error	True negative

Table 7.1 Confusion Matrix

In Equation 7.1, we count the number of true values ('True positive' and 'True negative'), and then in Equation 7.2 we calculated the RPCA.

By calculating the average value in the formula, half of the stocks are above the average in each round of prediction, and half of the stocks are below the average. If the groups are randomly sorted without any prediction, the RPCA calculated by the formula will be 0.5, so the benchmark for RPCA is always 0.5. Through this step, we eliminate the potential misleading results of the forecast arising from the overall trend of the market. For example, if we use 0 return as the benchmark to predict individual stock returns, then in a bull market, we only need to predict that all stocks will rise, then we will get a RPCA rate higher than 50%, but the construction principle of RPCA eliminates the market impact. As long as RPCA is higher than 0.5, our forecast for individual groups or are valid.

RPCA will be the main indicator to judge the validity of the model in the thesis.

7.5.3 Result of random forest

The results of year 2009 - year 2018 of Random Forest are shown in Figure 7.2:

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
RPCA	0.6	0.68	0.6	0.88	0.68	0.52	0.8	0.72	0.32	0.76

Table 7.2 The RPCAs of RFs on Chinese market forecasting task

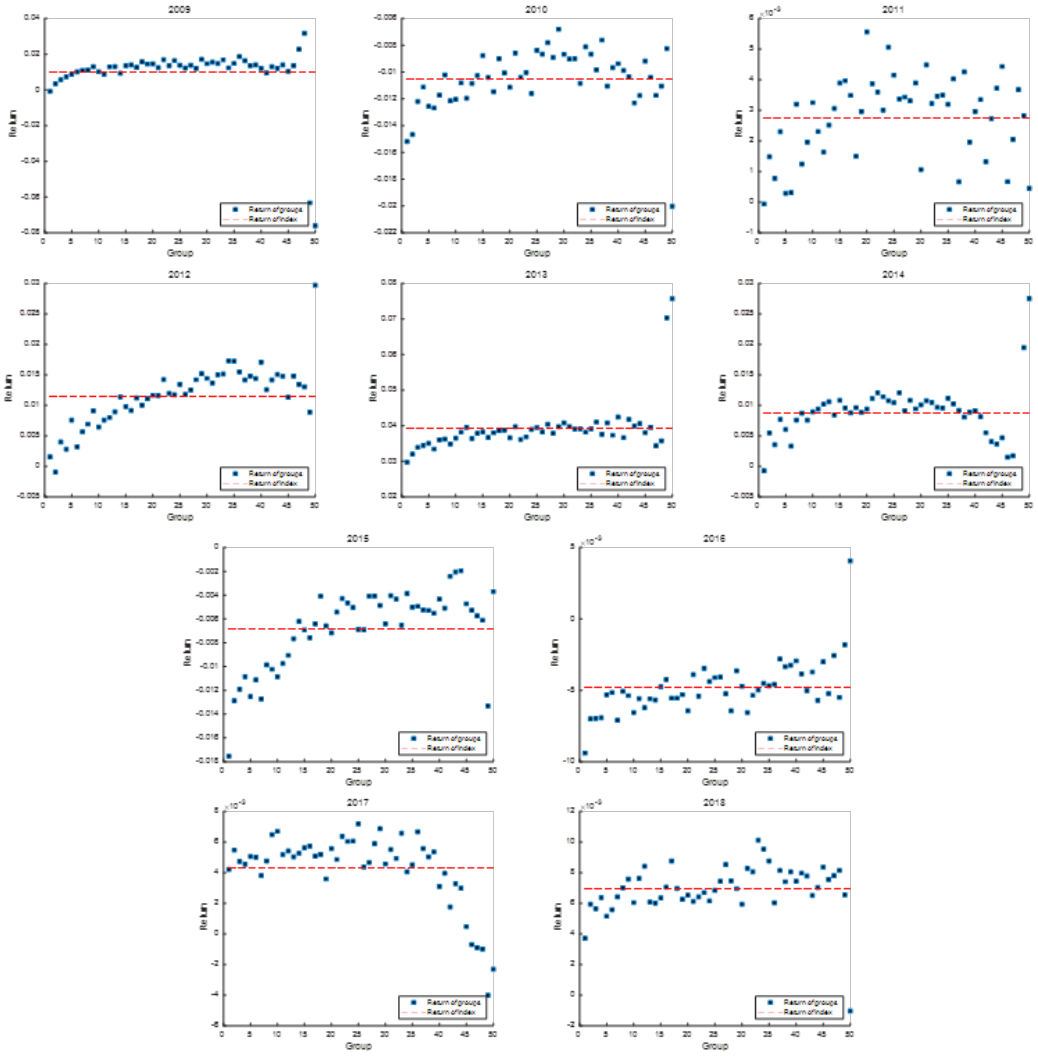


Fig. 7.2 Result of RFs, 2009 - 2018

The annual RPCA (RPCA; defined in section 7.5.2) of the random forest is shown in Table 7.2. From the table, we can conclude that the prediction of the years 2009, 2010, 2011, 2012, 2013, 2015, 2016, 2018 are effective. The average RPCA of all years is 0.656. The red line in the figure is the index return (mean of all groups) of the year.

As explained in Section 7.2, the result of the year 2014 actually shows the results of forecasting stock returns in the next year's period (2015.5.1-2016.5.1) using the 2014 stock financial indicators and the next year's stock technical indicators (2015.5.1-2016.5.1), and so on. In order to maintain the comparability of the results, the presentation of all the results of this study is annotated according to this method.

7.5.4 Result of Support Vector Machine

The results of the year 2009 - the year 2018 of Support Vector Machine are shown in Figure 7.3 and Table 7.3:

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
RPCA	0.84	0.96	0.76	0.96	0.92	0.80	0.92	0.88	0.64	0.92

Table 7.3 The RPCAs of SVMs on Chinese market forecasting task

We can conclude that all 10 years' results are effective from the table. The average RPCA of all years is 0.76.

7.5.5 Feature importance analysis

Whether it is SVMs and RFs, we can only understand how they operate from the architecture and principle, and they are black boxes seen from our sight in specific operations. However, we can judge the importance of each feature to the prediction result by the following method: change the value of a feature to a random number column and then measure the degree of reduction in the RPCA of the prediction. The larger the value, the greater the importance of the feature. This method is called the 'feature disturbance mean decrease effectiveness method'. We will use this method to rank the importance of feature input, select essential features to make predictions and view the results.

Using random forest, we calculated the average mean decrease in RPCA of 2014. The specific method calculates and records the mean decrease in RPCA every day after the forecast ends and averages the values throughout the year at last.

It can be concluded from Figure 7.4 that the importance of each feature is indeed different. We have 24 features in the feature space (9 technical and 15 financial indicators). The top 8

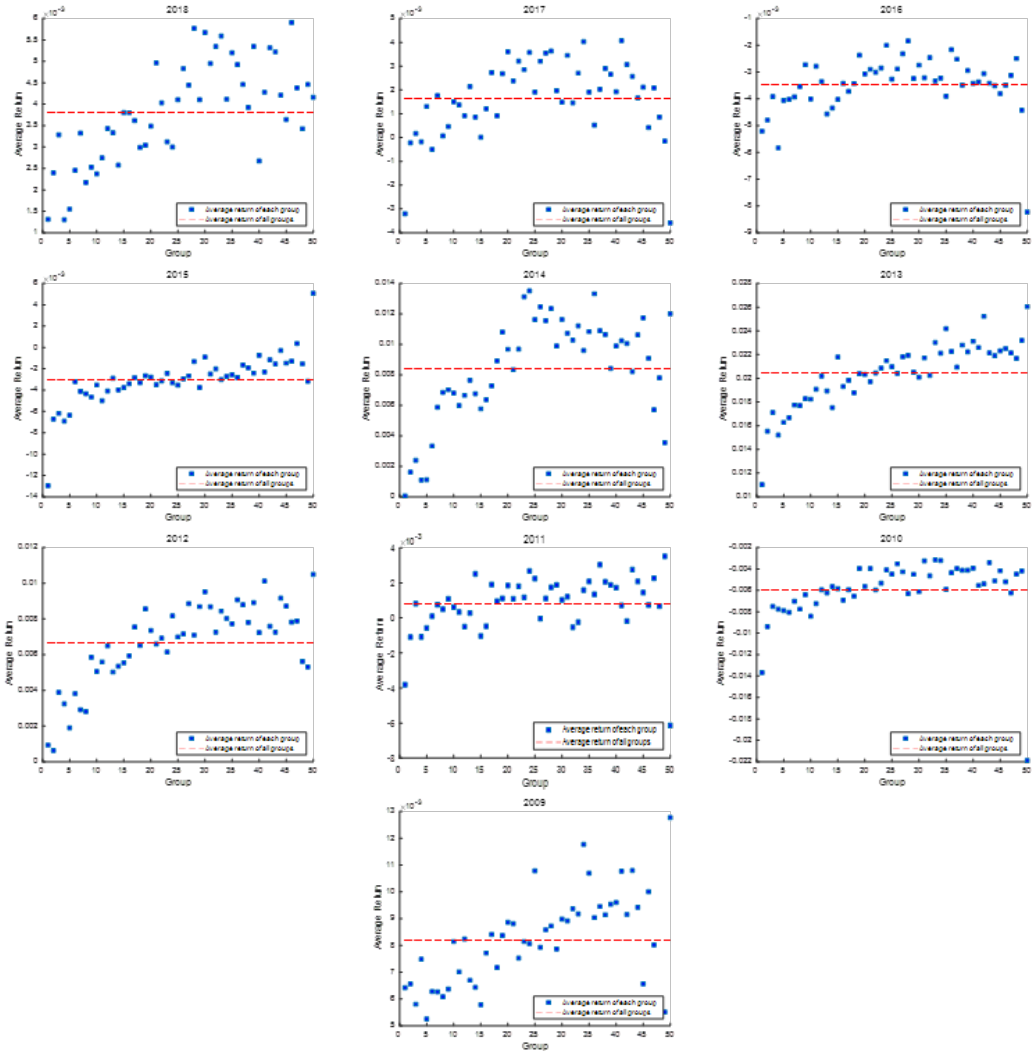


Fig. 7.3 Result of SVMs, 2009 - 2018

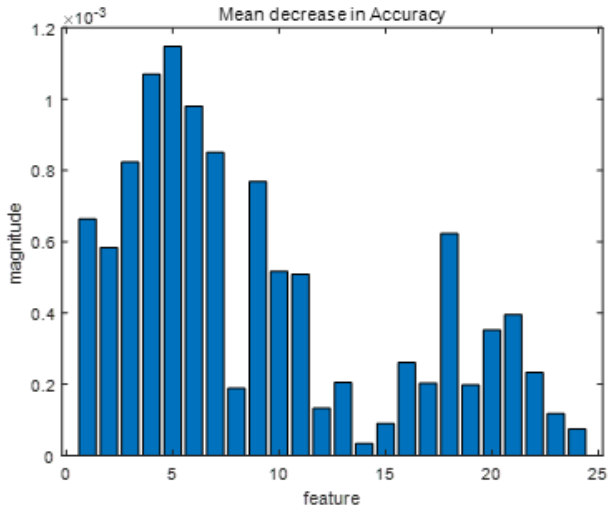


Fig. 7.4 Feature importance analysis

indicators of importance ranking are 5. MACD, 4. The stochastic oscillator, 6. MACD signal, 7. PROC, 3. RSI, 18. Net profit growth rate, 1. Long term relative position index, 2. Trading volume. The last 8 indicators are 14. Appreciation rate of capital preservation, 24. Working capital turnover 15. Fixed assets growth rate 23. Inventory turnover 12. Return on equity 8. OBV, 13. The growth rate on main business income,17. Undistributed profit per share.

In Figure 7.5 we use the top and the bottom 8 indicators as input features and contrast the result. The applying model is RFs.

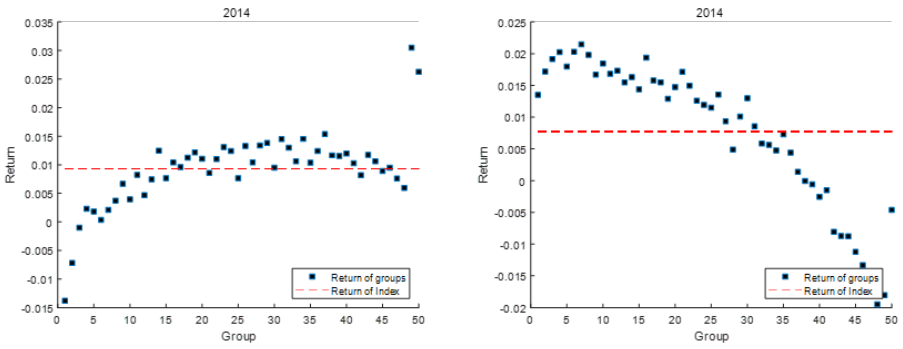


Fig. 7.5 Top and bottom 8 indicators result with RFs

To our surprise, using only the top 8 indicators as input, the RPCA of the prediction of 2014 is 0.68, which is even higher than the RPCA of the complete input (0.52), while as Figure 7.5 shows, the RPCA of using the bottom 8 importance acquired is only 0.08. The result proves that even if some indicators of lower importance are deleted, the prediction is still effective (even more effective).

It implies that low-importance indicators may have a zero or even negative impact on our prediction results, considering the efficiency, it might be worthy to delete those indicators. The highly important features are primarily technical analysis indicators, and the stocks with low importance are mostly financial indicators. In Section 7.6, we will test the prediction performance of just using technical indicators.

7.6 Result of Two Models with Technical Features

7.6.1 Result of random forest with technical features

The results of the year 2009 - the year 2018 of Random Forest using technical indicators are shown in Table 7.4 and Figure 7.6:

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
RPCA	0.84	0.96	0.76	0.96	0.92	0.8	0.92	0.88	0.64	0.92

Table 7.4 The RPCA of RFs with technical feature

We can conclude that all years' results are effective from the table. The average RPCA is 0.86.

7.6.2 Result of support vector machine with technical features

The results of the year 2009 - the year 2018 of Support Vector Machine are shown in Table 7.7 and Figure 7.5:

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
RPCA	0.68	0.64	0.68	0.8	0.84	0.68	0.8	0.68	0.56	0.8

Table 7.5 The RPCA of SVMs with technical feature

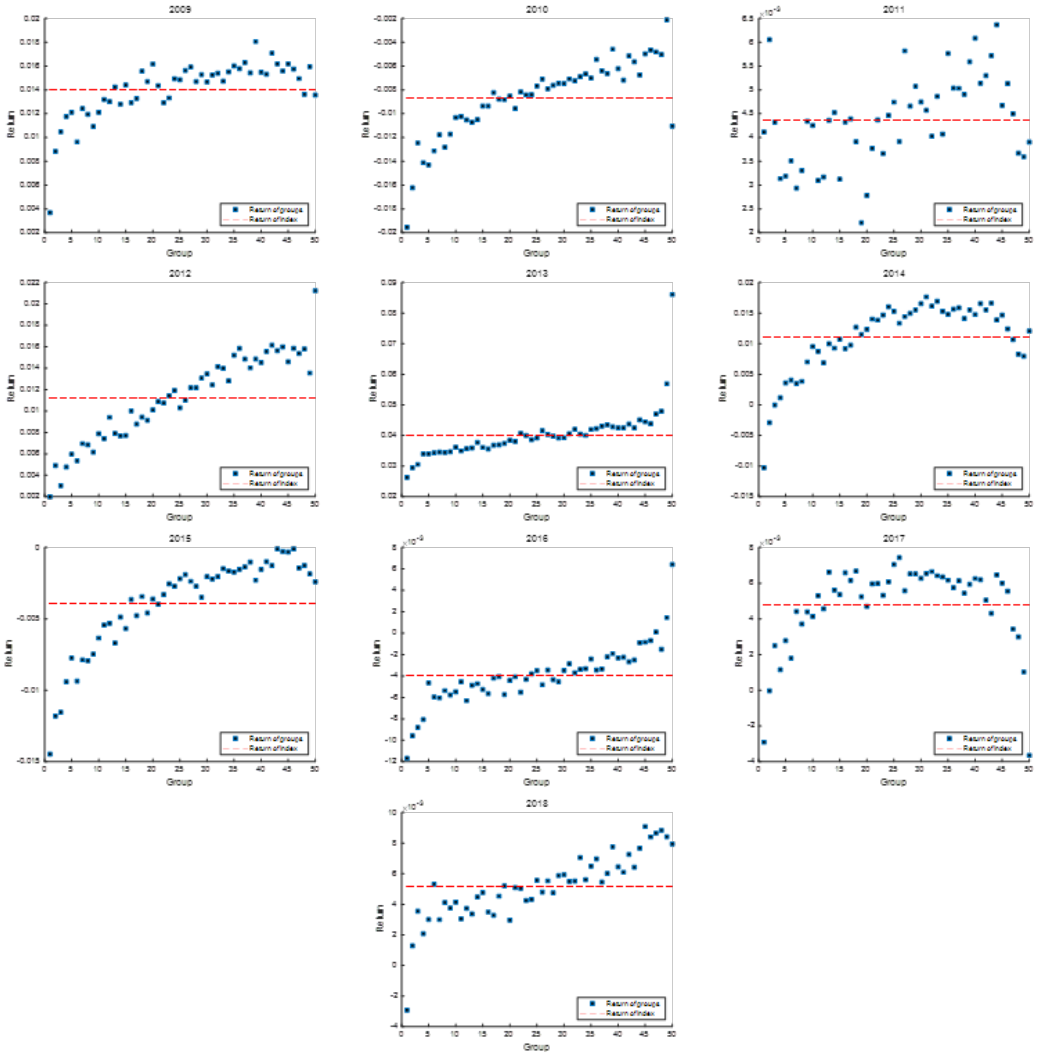


Fig. 7.6 Result of RFs with technical feature

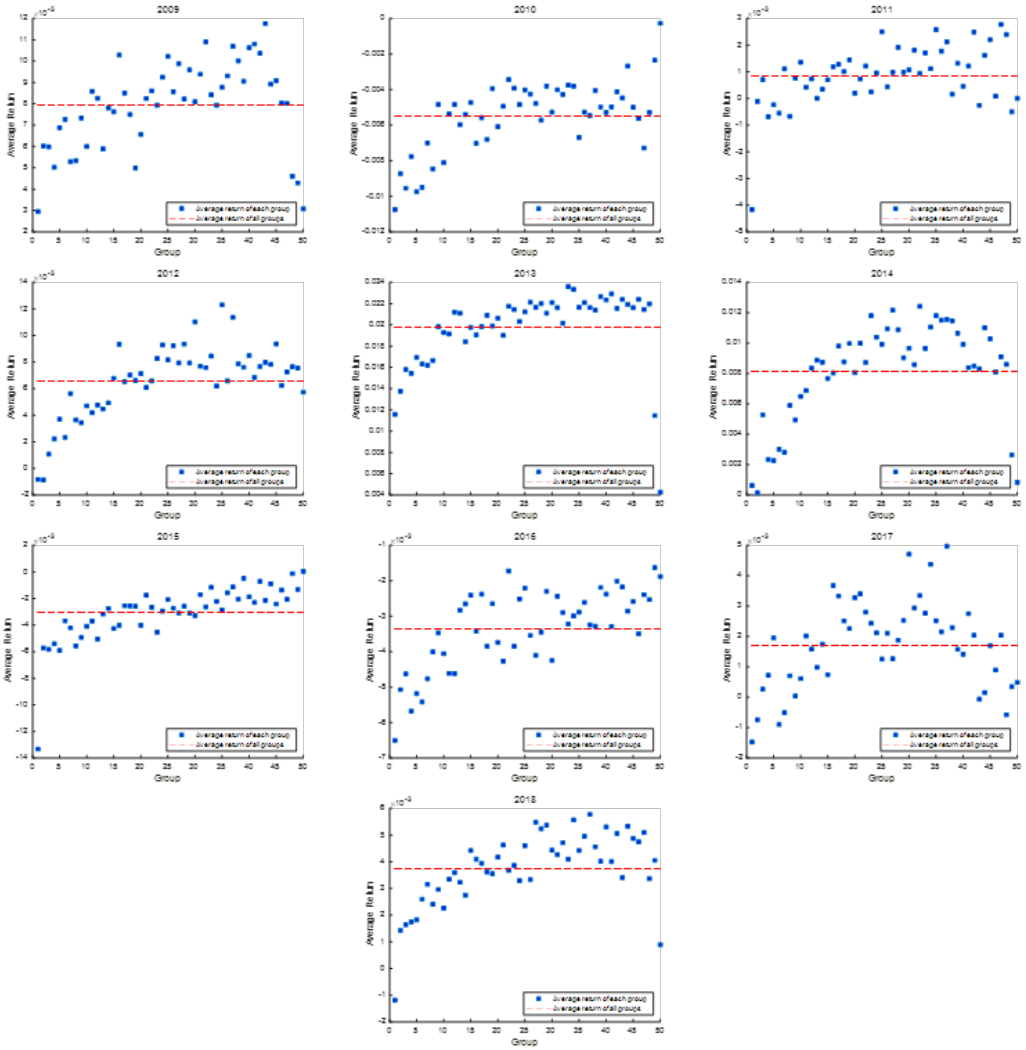


Fig. 7.7 Result of SVMs with technical feature

We can conclude that except for the year 2017, the other nine years' results are effective, especially in 2012, 2013, 2015 and 2018. The average RPCA of all years is 0.716, as shown in Table 7.5.

7.7 SVMs vs RFs, Technical Features vs Financial Features

RFs	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**Ave**	**Row**
All	0.6	0.68	0.6	0.88	0.68	**0.52**	0.8	0.72	**0.32**	0.76	0.66	1
Technical	0.84	0.96	0.76	0.96	0.92	0.8	0.92	0.88	0.64	0.92	0.86	2
Financial	**0.4**	**0.36**	**0.32**	0.6	**0.32**	**0.08**	**0.48**	**0.48**	**0.16**	**0.32**	0.35	3
SVMs	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**Ave**	
All	0.76	0.8	0.64	0.84	0.84	0.76	0.84	0.6	0.64	0.88	0.76	4
Technical	0.68	0.64	0.68	0.8	0.84	0.68	0.8	0.68	0.56	0.8	0.72	5
Financial	0.52	0.76	0.6	0.88	**0.44**	**0.12**	0.68	0.8	0.64	0.56	0.6	6

Table 7.6 RPCA comparing SVMs and RFs distinguishing technical and financial features

Besides testing using all of the features in Section 7.5 and only technical features in Section 7.6, we tested the data independently using financial features, and the competed result is shown in Table 7.6. For each model (RFs and SVMs), there are three lines of results generated using different input features, 'All' means to use all of the indicators, 'Technical' means only use technical features, 'Financial' means only use financial indicators. The box filled with thicker indicates that the forecast failed for that year. From 7.6, we can draw the following analysis:

1. When using all indicators and technical indicators, both algorithms are practical. The effectiveness of RFs on technical indicators ooutperformed SVMs, while the effectiveness of SVMs on all indicators outperformed RFs. When using only financial indicators, only SVMs could generate weak effectiveness. The RFs with technical inputs has the most potent prediction ability.

2. Compare the result of row 2 and row 3, row 5 and row 6, we can conclude that the predicting effectiveness of technical features is better than the financial features. At the same time, there is a large gap between the results of RFs (0.86 vs 0.35) and a small gap between the results of SVMs (0.72 vs 0.6).

3. Analyse rows 1, 2 and 3, the RFs makes reverse prediction when using financial indicators (0.35) while making robust, compelling predictions (0.86) when using

technical indicators, then it is imaginable that the RFs achieved a median RPCA of 0.66 after combining the two kinds of indicators as inputs.

From rows 4, 5 and 6, we conclude that the SVMs makes weak positive predictions when using financial indicators (0.6) and makes effective predictions (0.72) when using technical indicators. After combining both kinds of indicators, it achieves enhanced effectiveness of 0.76 RPCA.

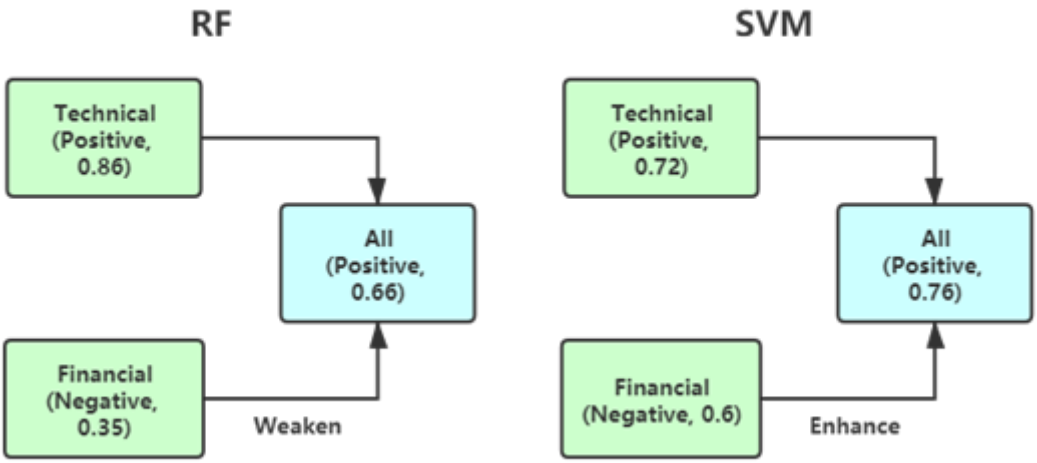


Fig. 7.8 Intrinsic technical route of RFs and SVMs

As Figure 7.8 shows, this revealed the different intrinsic technical routes of RFs and SVMs. Based on the construction principle of the random forest, it is understandable that when combining features with effectiveness and reverse effectiveness, the random forest will give out a final result with median effectiveness between those two kinds of features. Similarly, it seems that the support vector machine would generate an enhanced effectiveness result after combining features with strong effectiveness and weak effectiveness.

Hence the unsolved confusion is: when applying the same financial indicators as input, why do SVMs and RFs give opposite results (One is weakly positive and the other is strongly reversed)?

Financial indicators are undoubtedly meaningful to predicting the return of stocks because financial indicators carry certain information (if the data set is the noise that does not carry information, the prediction RPCA will be close to 0.5). One hypothesis is that the difference in performance may be related to the importance of the financial and technical ratios since most of the financial ratios are unimportant. If this hypothesis is proved, then the SVMs

may have a stronger regression ability on the unimportant (still need to carry information) features.

7.8 Robustness Test

From the analysis in Section 7.7, we can see that random forest combined with technical indicators has the best predictive effect. Therefore, in the following part of Chapters 7 and 8, unless otherwise specified, we will use RFs with technical indicators as the primary test model. In Section 7.7, we proved that discarding unimportant features have no negative impact on the prediction. We will run further robustness tests in this section.

In subsection 7.8.1, we checked the prediction effectiveness under bull and bear market condition. In subsection 7.8.2, we used different forecast horizons (the length of the rolling window) to test the algorithm's robustness. In subsection 7.8.3 the optimisation effectiveness of SVMs is tested.

7.8.1 Bull and Bear market

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Average
Index Return	0.014	-0.009	0.004	0.011	0.04	0.011	-0.004	-0.004	0.005	0.005	0.007
Up days Ratio	0.645	0.382	0.523	0.6	0.801	0.579	0.533	0.478	0.54	0.537	0.562
RPCA (RFs - T)	0.84	0.96	0.76	0.96	0.92	0.8	0.92	0.88	0.64	0.92	0.86

Table 7.7 RPCA under different market condition

In Table 7.7, the first row of the table is the average return of all stocks in the corresponding period (referred to as market index in this thesis), which is consistent with the previous result. The second row is the ratio of market up days (days when market return is greater than 0) to total trading days of the period. The third row is the RPCA of random forest model with technical ratios.

'Index Return' and 'Up days Ratio' shows consistency of rising and falling. The RFs-T (Random Forest using technical features) consistently get a high RPCA through the whole data period no matter the market condition. We can conclude that our prediction results are effective in both bull and bear markets.

Days	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	Average
5	0.8	0.92	0.76	0.72	0.96	0.92	0.92	0.76	0.84	0.84	0.844
10	0.84	0.96	0.76	0.96	0.92	0.8	0.92	0.88	0.64	0.92	0.86
15	0.8	0.76	0.96	0.96	0.96	0.68	0.96	0.8	0.52	0.88	0.828
20	0.6	0.6	0.88	0.96	0.96	0.72	0.92	0.52	0.72	0.24	0.712
40	0.92	0.84	0.44	0.84	0.72	0.92	1	0.28	0.76	0.28	0.7
60	0.96	0.88	0.48	0.92	0.8	0.88	0.96	0.12	0.84	0.76	0.76

Table 7.8 Robustness test - forecast horizon change

7.8.2 Change forecast horizon

We applied different forecast horizons (width of the rolling window) in Table 7.8. The average RPCA of the 6 forecast horizons are beyond 50%, which shows that the model is not sensitive to forecast horizon parameter changes.

However, when the width of the rolling window is set to 5, 10, and 15, the effectiveness is better than the value of 20, 40, and 60. This may be related to the fact that our technical indicators are mainly short-term indicators.

7.8.3 Optimisation effectiveness of SVMs

As mentioned earlier, we applied genetic algorithm to optimise the three parameters: C, σ, ε, in Matlab GA built-in function, the corresponding parameters name is c, g, p, the optimisation bound is: $c : [0.1, 10], g : [0.1, 10], p : [0.01, 1]$.

How to prove that our optimisation is effective? The most direct method is to compare the results produced by the optimised parameters with the results produced by the parameters that have not been optimised.

We take the boundary value of the optimisation range of the three parameters as the control group. If the predicted result of the control group fails to exceed the optimised result, it proves that our optimisation is effective.

The RPCA of the control group of each year and the average is shown in Table 7.9. While the average RPCA of optimised SVMs with technical indicators of all years is 0.716, from the table, we can conclude that 5 out of 8 control results are far below the optimised result, 3 out of 8 results are slightly higher than the optimised result. The result shows that although prior empirical parameters settings could generate positive result on the certain data, our GA optimisation is efficient.

In addition, it can assist in judging whether the optimisation is effective from the distribution of the optimised parameters: an effective optimisation will inevitably make the optimised parameters concentrate in a small space of the search space. If the optimised

	c=10, g=10, p=1	c=10, g=0.1, p=1	c=0.1, g=10, p=1	c=0.1, g=0.1, p=0.01
2009	0.48	0.48	0.44	0.76
2010	0.44	0.44	0.6	0.84
2011	0.6	0.6	0.48	0.56
2012	0.72	0.72	0.8	0.8
2013	0.28	0.28	0.76	0.72
2014	0.64	0.64	0.6	0.72
2015	0.28	0.28	0.52	0.88
2016	0.56	0.56	0.72	0.72
2017	0.32	0.32	0.52	0.64
2018	0.68	0.68	0.52	0.64
Average	**0.5**	**0.5**	**0.596**	**0.728**
	c=10, g=10, p=0.01	c=10,g=0.1,p=0.01	c=0.1, g=10, p=0.01	c=0.1, g=0.1, p=1
2009	0.8	0.72	0.68	0.76
2010	0.56	0.88	0.64	0.84
2011	0.4	0.64	0.64	0.56
2012	0.68	0.8	0.84	0.8
2013	0.6	0.76	0.64	0.72
2014	0.56	0.68	0.64	0.72
2015	0.68	0.88	0.76	0.88
2016	0.36	0.72	0.44	0.72
2017	0.48	0.6	0.6	0.64
2018	0.48	0.8	0.64	0.64
Average	**0.56**	**0.748**	**0.652**	**0.728**

Table 7.9 The RPCA of control parameters groups

results are randomly distributed in the search space (In other words, there is no difference between optimising parameters and randomly selecting parameters), then optimisation is not adequate.

However, it should be noted that the above inference cannot be reversed. That is, just observing the parameter concentration does not prove that the optimisation is effective. Therefore, parameter concentration is a necessary and insufficient condition for effective optimisation.

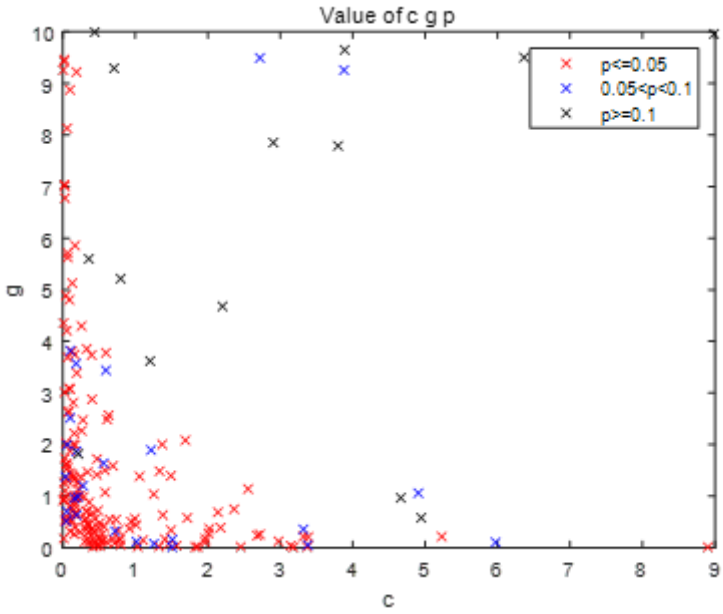


Fig. 7.9 Value of three parameters

We plot each day's parameters value after the optimisation of the year 2014. In Figure 7.9, the $X-axis$ is the value of C, the Y-axis is the value of g, and three different colors are used to represent the value of p (In white/black printing version p can not be identified). From Figure 7.9 we can roughly conclude that the most optimised C and g distribute in [1, 3], and most p are under 0.05. Figure 7.10 tells the same story from the distribution plot.

Based on the above analysis from two aspects, GA optimisation is effective.

7.9 Trading Simulation

In order to evaluate the potential application of the research, we write a program to simulate the stock trading based on our prediction. The matlab code are in Algorithm 7.2.

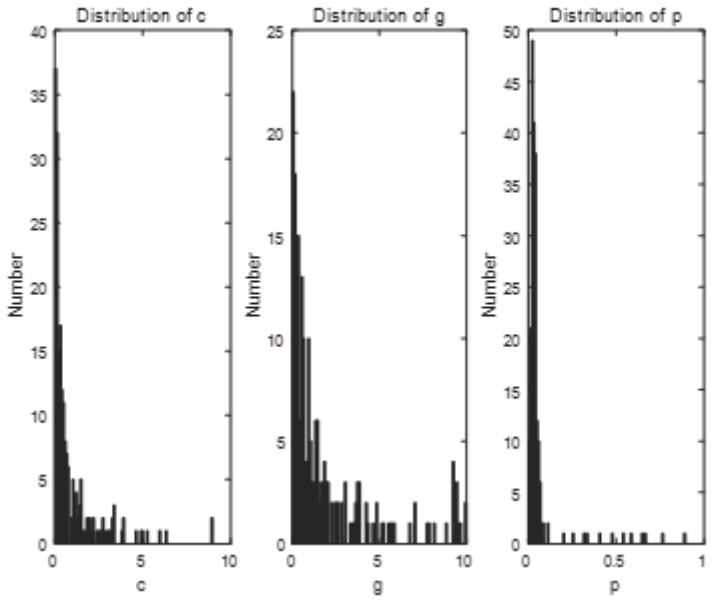


Fig. 7.10 Distribution of the three parameters

7.9.1 Simulation without trading issues

Firstly we do not consider transaction issues and broker fees and simulate the trading process to glance at the result.

1. Investment targets

   We select the top 5 groups of stocks (250 stocks) as investment targets. Investment capital was evenly distributed to each stock.

2. Position adjusting frequency

   Our forecast step length is 10 days, so the trading position adjustment step length is also 10 days. Every 10 trading days, the position will be adjusted according to the forecast result of the day, and the position will be held in the 10 trading days between the two transactions.

3. Benchmark

   We calculate the mean of the average return of all stocks in that year as the index benchmark.

Algorithm 7.2 Trading simulation

```
Allo=500; % How many stocks the capital would be invested in
% signal & return
a = test_y M;
idx1 = find(isnan(a(3,:)) == 1);
a(:, idx1) = [];
C = [];
for i=1:size(a,2) do
    b = a(:, i);
    b(isnan(b)) = [];
    c = flipud(b);
    C = [C, c(1 : Allo)];
end for
C1 = C + 1;
C2 = [C2, C1];
RMAHA = C2(1 : Allo, 1 : (argo − 1) : end);
TSEE = prod(RMAHA, 2);
```

The denoted trading period is from 2009 to 2018, the exact trading period settings are explained in Section 7.2. During the period, the return of index is:

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Benchmark	0.014	-0.009	0.004	0.011	0.04	0.011	-0.004	-0.004	0.005	0.005

Table 7.10 Index return of year 2009-2018

The return of predicted top 5 groups is:

Benchmark	G1	G2	G3	G4	G5
0.0716	6.26	3.73	3.39	6.55	2.41

Table 7.11 The return of top 5 groups without transaction fee

From Table 7.11, we can conclude that every group achieved excess return over the benchmark, but this result is a simple preliminary result that does not consider the transaction fees and other issues in the actual trading.

7.9.2 Simulation with trading issues

When come to real trading, we must consider transaction fees and other real issues:

1. Transaction fee

   In the Chinese stock exchange, transaction fees are mainly composed of stamp duty 1‰ (collected when sell) and brokerage commission 0.1‰ - 3‰ (depending on the broker's policy and investor's capital, collected when sell and buy), typical commission for small investors are 0.3 ‰. Thus transaction fee would be 0.3‰ *2 + 1‰ =1.6‰.

2. The GEM stocks

   The Chinese Growth Enterprise Market, also known as the Second-board Market, is a type of securities market different from the Main-Board Market. It is designed for entrepreneurial companies that are temporarily unable to be listed on the Main-Board Market, similar to the Nasdaq stock exchange in the United States.

   There are capital and trading qualification restrictions for trading GEM stocks so we will exclude GEM stocks from the optional investment target in the trading simulation.

3. Exclude stocks that reach daily limit with opening price

   The Chinese stock market implements a price limit system. The price limit system stipulates that the maximum fluctuation range of the stock price in one trading day is 10% up and down from the closing price of the previous trading day, and the transaction is restricted after the stock reaches the point. The restriction mechanism is that the bid price shall not exceed the daily limit price.

Before the free trading opening in the Chinese stock market (9:30 a.m.), there is a call auction for 5 minutes, and the call auction determines the stock's opening price. If the opening price is the daily limit, this phenomenon is called the daily limit opening.

For example, the stock's closing price on day 1 is 10 yuan, and the opening price on day 2 is 11 yuan, then it hits the daily limit, and any bid price cannot exceed 11 yuan.

In this situation, investors can buy stock A under two conditions: 1. Bid 11 yuan and queue up with other buyers with the same bid price, and there are usually many buying orders on this price, which makes the transaction almost impossible to complete. 2. If the price falls below the daily limit, investors can buy it instantly according to general stock trading rules.

To sum up, stocks that open at the daily limit price may not be available to be bought in actual trading, so we do not consider trading such stocks in the adjusted simulation.

Matlab code in Algorithm 7.3 are the algorithm of trading simulaition with trading issues.

Algorithm 7.3 Trading simulation with trading issues

```
tradetdx=tdx(ttr+rtlag-1:length(tdx));
tradenum=Allo;
signalmat=zeros(tradenum,length(tradetdx)-1);
for i=1:length(tradetdx)-1 do
    tempocst=Cst(:,i);
    for j=1:tradenum do
        ia=find(TCAtech(:,1)==tempocst(j) & TCAtech(:,2)==tradetdx(i+1));
        Tempoone=TCAtech(ia,:);
        if isempty(ia)==1 then
            signalmat(j,i)=1;
            % shilter 1:price limit up no buy chance
        else if Tempoone(3)==Tempoone(6) then
            signalmat(j,i)=1;
            %shilter2: GEM stocks
        else if Tempoone(1)>300000 & Tempoone(1)<600000 then
            signalmat(j,i)=1;
        end if
    end for
end for
```

Benchmark	Group 1	Group 2	Group 3	Group 4	Group 5	Broker fee
0.0716	4.56	2.62	2.36	4.79	1.61	0.1‰
0.0716	4.09	2.31	2.07	4.3	1.39	0.3‰
0.0716	3.66	2.03	1.81	3.85	1.18	0.5‰
0.0716	2.73	1.43	1.25	2.89	0.75	1‰
0.0716	0.53	-0.006	-0.08	0.6	-0.28	3‰

Table 7.12 The return of top 5 groups with different transaction fees

The result in Table 7.12 is generated by the same strategy running in the same period from 2009 to 2018 in Section 7.9.1: Every 10 days, the algorithm selects the highest ranked 5 groups of stocks (250 stocks) as investment targets and adjust the holding position, investment capital is evenly distributed to each stock. e.g., The Group 1 states top 1 group's result (containing 50 stocks), the 4.56 is the absolute return under 0.1‰ level broker fee, that means if investor invested 1 $ at the start he will get 5.56 $ at the end, while he can only get 1.0716 $ if he invested in the benchmark.

The adjusted simulation considered transaction fees excluded GEM stocks and stocks that reach the daily limit with the opening price. As explained in 7.9.2, in the Chinese stock exchange, transaction fees are mainly composed of stamp duty 1‰ (collected when sell) and

broker fee 0.1‰ - 3‰ (depending on the broker's policy and investor's capital, collected when sell and buy). The stamp duty is fixed, thus in the 'Broker fee' column in Table 7.12, when the broker fee is 0.1‰, the total transaction fee (including stamp duty) would be 0.1‰ *2 + 1‰ =1.2‰.

It can be concluded from the table that in addition to the highest brokerage commission rate of 3‰, when other rates are adopted, the stock returns of the top five groups of ten years are all higher than the benchmark, and the strategy can obtain excess returns. Since the strategy adjusts the position every ten trading days, it needs to adjust it 20 times a year. Under this frequency of transactions, the commission rate significantly impacts the final excess return.

Chapter 8

Result on HK and US Stock Market

8.1 Introduction

Through the preceding work of this research, we have established an effective machine learning model for stock return forecasting through the test on the stocks in Shanghai and Shenzhen stock exchanges in China, and we conclude that the most effective model and feature combination is Random Forest Regression with technical indicators. In this part, we will run the same RFs model on the stocks in HongKong, China (HK) stock market and United States (US) stock market for two research purposes:

1. To see how well the system generalizes its behaviour on other markets.
2. To compare the results, and note the differences of different markets.

The results on these two new markets are presented below.

8.2 Result on HK Market

The data of the HK market contains all the listed stocks of Hong Kong Exchanges and Clearing Limited (HKEX) from the year 2013 to 2018. The result are shown in Table 8.1 and Figure 8.1.

The RPCA(return prediction classification accuracy) of Hong Kong market cannot reach 50% for two years (the year 2014 and 2018), and the average RPCA is only 0.59.

Year	2013	2014	2015	2016	2017	2018	Average
	0.72	0.28	0.76	0.56	0.72	0.48	0.59

Table 8.1 Result on HK market

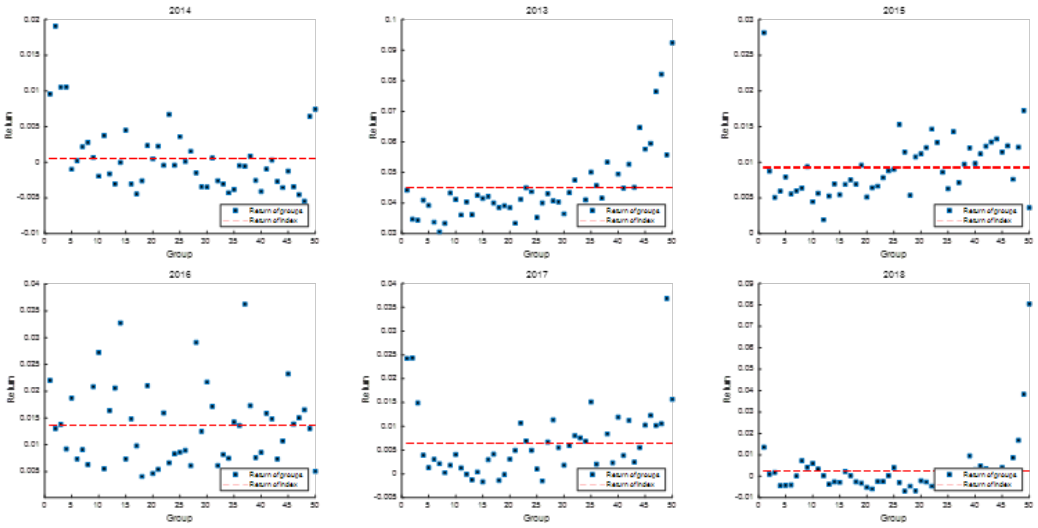


Fig. 8.1 Result on HK market

8.3 Result on US Market

The US market data contains the 500 most famous listed stocks of the New York Stock Exchange (NYSE) and the National Association of Securities Dealers Automated Quotation (NASDAQ) from the year 2013 to the year 2018. The result is as follows:

Year	2013	2014	2015	2016	2017	2018	Average
	0.84	0.94	0.81	0.9	0.92	0.87	0.91

Table 8.2 Result on US market

The result is good and the RPCA exceed 0.8 in every year. The average RPCA of the US market is 0.91, it is slightly better than the Chinese market (RPCA: 0.85).

8.4 Result Analysis

8.4.1 Effectiveness

The results show that the RPCA of the US market (RPCA: 0.91) is slightly better than the Chinese market (RPCA: 0.85). The Hong Kong market cannot reach 50% for two years (the year 2014 and 2018), and the average RPCA is only 0.59, which is significantly lower than

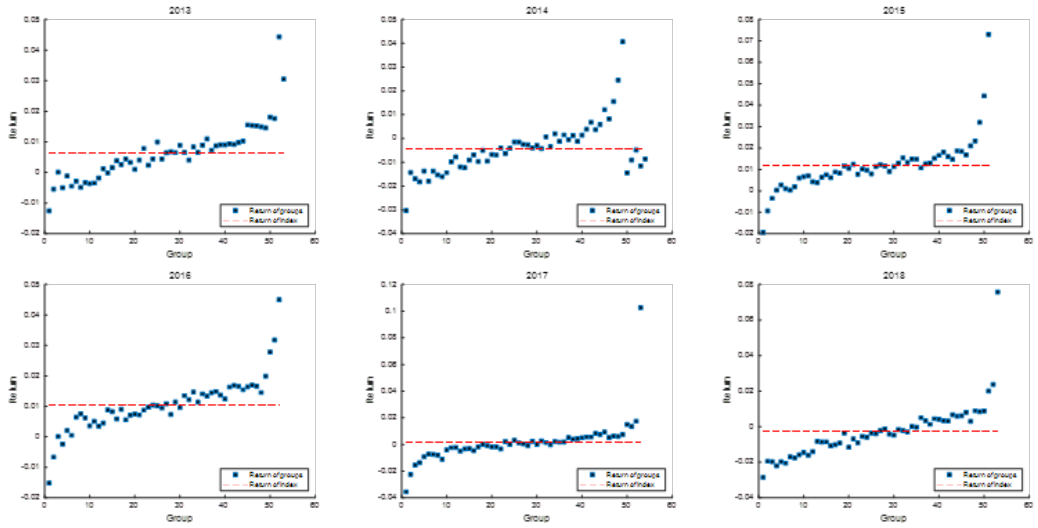


Fig. 8.2 Result on US market

Year	2013	2014	2015	2016	2017	2018	Average
HK	0.72	0.28	0.76	0.56	0.72	0.48	0.59
US	0.84	0.94	0.81	0.9	0.92	0.87	0.91
Mainland China	0.92	0.8	0.92	0.88	0.64	0.92	0.85

Table 8.3 Comparable return prediction classification accuracy of the US, HK and China markets

the other two markets. We can conclude that the model is significantly effective in the China and US markets but somewhat less effective in the Hong Kong market.

8.4.2 Result analysis from the view of Efficient Market Hypothesis

The input of the forecasting model is technical indicators. We can compare the results by inputting data from different country markets, the significant difference between the results produced by the same input data will be an intuitive manifestation of market efficiency.

8.5 Efficient Market Hypothesis

The efficient market hypothesis believes that in a stock market with strict laws, good functions, high transparency, and sufficient competition, all valuable information has been reflected in the stock price, including the current and future value of the company unless it exists market manipulation, it is impossible for investors to obtain excess profits above the market average by analysing historical information.

The efficient capital market hypothesis has three progressive forms [31]:

1. Weak-Form Market Efficiency

   This hypothesis states that in the case of weak-form validity, the market price has fully reflected all historical security price information, including stock trading prices, trading volumes,etc.

   Corollary: If the weak-form efficient market hypothesis is achieved, the technical analysis of stock prices will be useless, and fundamental analysis may still help investors obtain excess profits.

2. Semi-Strong-Form Market Efficiency

   The hypothesis believes that the price has fully reflected all publicly available information about the company's fundamental prospects.

   Corollary: If the semi-strong effective hypothesis is established, the use of fundamental analysis in the market will be useless, but inside information may still obtain excess profits.

3. Strong-Form Market Efficiency

   The strong-form efficient market hypothesis believes that prices have fully reflected all the information about the company's operations, including information that has been disclosed or not.

   Corollary: In a strong-form efficient market, no investors obtain excess profits.

8.5.1 Bounded Rationality

The basic premise of the efficient market hypothesis is the rational person hypothesis. That is, the analytical ability of investors is homogeneous, and every investor can make rational inferences based on the information they have [32].

A large number of studies in behavioral finance have proved that this assumption is not steady. Investors only have bounded rationality, which is limited by everyone's analytical ability and available analytical resources [32]. For example, although the machine learning model used in this study is still statistical learning in nature, it is quite different from the simple statistical method used in indicator analysis 100 years ago. We do not intend to go deep into academic debates, nor do we intend to accurately define whether a particular market has reached a certain degree of efficiency. However, if we use the same indicator and the same model to test the prediction results of different markets, it generates significantly different results. Then we can cautiously conclude: under this test, one market is more efficient than the other.

Therefore, we can conclude from the result that the Hong Kong stock market is more effective than the Chinese stock market and US stock market during our forecast period under this test.

Chapter 9

Conclusion

The research uses Fuzzy Clustering to cluster and select 15 financial indicators from 42 financial indicators each year, combine 8 technical indicators to construct the input space, and then construct GA-SVR and Random Forest to predict stock returns. We apply both statistical indicators and trading simulations to evaluate the results. The results show that GA-SVR and Random Forest can effectively predict the 10-day stock return. By applying different kinds of indicators, the algorithms would have different prediction abilities. The Random Forest with technical indicators has the highest prediction effectiveness.

We achieve the research purpose. The main contribution of the research are:

1. The research systematically quantified the financial ratios of the Chinese stock market stocks from 2009 to 2018. Compared with other research that only selected a short period, this research tested ten years of data, including a complete bear market and bull market cycle, making the conclusions more steady.

2. The research proved that the stock's 10-days relative return could be predicted. The Random forest regression with technical indicators has the best predictive ability.

3. With similar positive prediction results on combining data of two input features (high importance features vs. low importance features, or technical features vs. financial features), SVMs and RFs show different effectiveness on two kinds of features. RFs outperformed SVMs on high importance features, while SVMs outperformed RFs on low importance features.

4. We compared the prediction results of China, Hong Kong and US market from 2013 to 2018 by using RFs with technical indicators. The results show that in six years, the prediction RPCA of the China and US market exceeds 50% every year, while that of the Hong Kong market cannot reach 50% for two years. The average forecast RPCA

of HK is only 0.59, which is significantly lower than the Chinese forecast RPCA of 0.85 and the US forecast RPCA of 0.91. Therefore, we can conclude that the Hong Kong stock market is more effective than the Chinese stock market and American stock market during our forecast period under this test.

5. This research proved that when forecasting time series data (including single time series data and panel data), the accuracy obtained by the following two methods is not accurate: 1. Use out-of-bag accuracy instead of actual test accuracy 2. For time-series data, divide the training and test sets based on the time axis.

There are areas worthy of improvement and further research:

1. There are phenomenon we can not explain. In Section 7.6, our two models have achieved high prediction effectiveness. However, by observing the results of 2011, 2014, 2015, 2017 of RFs, and 2009, 2013, 2014, 2017, 2018 of SVMs, we can find that in these years, although the forecast plots show an upward trend, they declined at tails. The reason for this phenomenon is still unclear. It can explain why the simulated trading return in Section 7.9 did not have a significant linear decline with the number of groups.

2. We found that SVMs and RFs show different effectiveness on two kinds of features. The reason behind this phenomenon needs further research.

References

[1] Luckyson Khaidem, Snehanshu Saha, and Sudeepa Roy Dey. Predicting the direction of stock market prices using random forest. *arXiv preprint arXiv:1605.00003*, 2016.

[2] Tian hao. Research on the trading strategy of shanghai and shenzhen 300 stock index futures based on xgboost. Master's thesis, Shanghai Normal University, 2018.

[3] Zou Y. *Research on Shanghai and Shenzhen 300 Index Trend Forecast Based on Machine Learning*. PhD thesis, Shandong University, 2018.

[4] Xianghui Yuan, Jin Yuan, Tianzhao Jiang, and Qurat Ul Ain. Integrated long-term stock selection models based on feature selection and machine learning algorithms for china stock market. *IEEE Access*, 8:22672–22685, 2020.

[5] XingYu Fu, JinHong Du, YiFeng Guo, MingWen Liu, Tao Dong, and XiuWen Duan. A machine learning framework for stock selection. *arXiv preprint arXiv:1806.01743*, 2018.

[6] Qasem A Al-Radaideh, Adel Abu Assaf, and Eman Alnagi. Predicting stock prices using data mining techniques. In *The International Arab Conference on Information Technology (ACIT'2013)*, pages 1–8, 2013.

[7] Zheng Tan, Ziqin Yan, and Guangwei Zhu. Stock selection with random forest: An exploitation of excess return in the chinese stock market. *Heliyon*, 5(8):e02310, 2019.

[8] Sahaj Singh Maini and K Govinda. Stock market prediction using data mining techniques. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pages 654–661. IEEE, 2017.

[9] Zhou Ming. Stock selection with random forest in chinese market. *Trade Communication*, 2018.

[10] KS Loke. Impact of financial ratios and technical analysis on stock price prediction using random forests. In *2017 International Conference on Computer and Drone Applications (IConDA)*, pages 38–42. IEEE, 2017.

[11] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

[12] James Foye. A comprehensive test of the fama-french five-factor model in emerging markets. *Emerging Markets Review*, 37:199–222, 2018.

[13] Kent L Womack and Ying Zhang. Understanding risk and return, the capm, and the fama-french three-factor model. *Available at SSRN 481881*, 2003.

[14] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.

[15] Anke Meyer-Baese, Axel Wismueller, and Oliver Lange. Comparison of two exploratory data analysis methods for fmri: unsupervised clustering versus independent component analysis. *IEEE Transactions on Information Technology in Biomedicine*, 8(3):387–398, 2004.

[16] R Baumgartner, L Ryner, W Richter, R Summers, M Jarmasz, and R Somorjai. Comparison of two exploratory data analysis methods for fmri: fuzzy clustering vs. principal component analysis. *Magnetic Resonance Imaging*, 18(1):89–94, 2000.

[17] Eddy Liu. Soft margin, 02 2018.

[18] Alfred Cowles 3rd. Can stock market forecasters forecast? *Econometrica: Journal of the Econometric Society*, pages 309–324, 1933.

[19] Eugene F Fama and James D MacBeth. Risk, return, and equilibrium: Empirical tests. *Journal of political economy*, 81(3):607–636, 1973.

[20] Eugene F Fama. Efficient capital markets: Ii. *The journal of finance*, 46(5):1575–1617, 1991.

[21] A Salski. Fuzzy clustering of fuzzy ecological data. *Ecological Informatics*, 2(3):262–269, 2007.

[22] Taeshik Shon, Yongdae Kim, Cheolwon Lee, and Jongsub Moon. A machine learning framework for network anomaly detection using svm and ga. In *Proceedings from the sixth annual IEEE SMC information assurance workshop*, pages 176–183. IEEE, 2005.

[23] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[24] Mariette Awad and Rahul Khanna. Support vector regression. In *Efficient learning machines*, pages 67–80. Springer, 2015.

[25] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[26] Yuan Ren and Guangchen Bai. Determination of optimal svm parameters by using ga/pso. *J. Comput.*, 5(8):1160–1168, 2010.

[27] Zheng Chunhong and Jiao Licheng. Automatic parameters selection for svm based on ga. In *Fifth world congress on intelligent control and automation (IEEE Cat. No. 04EX788)*, volume 2, pages 1869–1872. IEEE, 2004.

[28] SU Zhi and FU Xiaoyuan. Kernel principal component genetic algorithm and improved svr stock selection model. *Statistical Research*, 30(5):54–62, 2013.

[29] Ching-Hsue Cheng and Huei-Yuan Shiu. A novel ga-svr time series model based on selected indicators method for forecasting stock price. In *2014 International Conference on Information Science, Electronics and Electrical Engineering*, volume 1, pages 395–399. IEEE, 2014.

[30] MJ Valadan Zoej, Mehdi Mokhtarzade, Ali Mansourian, Hamid Ebadi, and Saeid Sadeghian. Rational function optimization using genetic algorithms. *International journal of applied earth observation and geoinformation*, 9(4):403–413, 2007.

[31] Sanjoy Basu. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The journal of Finance*, 32(3):663–682, 1977.

[32] Robert J Aumann. Rationality and bounded rationality. In *Cooperation: Game-Theoretic Approaches*, pages 219–231. Springer, 1997.

Appendix A

Fuzzy Clustering Result

A.1 Construct Fuzzy Similarity Matrix

The financial ratios of profitability, development capability, shareholders' equity (probability), solvency, and operating capability of the listed company are used to construct the fuzzy similarity matrix as the Table A.1, Table A.2, Table A.3, Table A.4, Table A.5 show, this demonstrated result example is from the year 2014.

The Fuzzy Similarity Matrix of Profitability Ratios									
	F1-1	F1-2	F1-3	F1-4	F1-5	F1-6	F1-7	F1-8	F1-9
F1-1	1	0.1299	0.0338	0.0377	0.0236	0.016	0.9134	0.0043	0.0474
F1-2	0.1299	1	0.8158	0.8179	0.8196	0.2042	0.3597	0.0261	0.0326
F1-3	0.0338	0.8158	1	0.9992	0.9904	0.1565	0.1586	0.0622	0.0355
F1-4	0.0377	0.8179	0.9992	1	0.9893	0.1587	0.1586	0.0635	0.0357
F1-5	0.0236	0.8196	0.9904	0.9893	1	0.1452	0.1618	0.0571	0.0149
F1-6	0.016	0.2042	0.1565	0.1587	0.1452	1	0.0278	0.0153	0.0109
F1-7	0.9134	0.3597	0.1586	0.1586	0.1618	0.0278	1	0.006	0.0244
F1-8	0.0043	0.0261	0.0622	0.0635	0.0571	0.0153	0.006	1	0.0121
F1-9	0.0474	0.0326	0.0355	0.0357	0.0149	0.0109	0.0244	0.0121	1

F1-1 Operating gross profit margin
F1-2 Operating income net profit margin
F1-3 Return on assets
F1-4 Net profit rate of total assets
F1-5 Return on current assets
F1-6 Return on fixed assets
F1-7 Profit margin
F1-8 Return on equity
F1-9 Growth rate of main business income

Table A.1 The fuzzy similarity matrix of profitability ratios

The Fuzzy Similarity Matrix of Development Capability Ratios					
	F2-1	F2-2	F2-3	F2-4	F2-5
F2-1	1	1	0.0043	0.6936	0.0216
F2-2	1	1	0.0043	0.6936	0.0216
F2-3	0.0043	0.0043	1	0.0506	0.0988
F2-4	0.6936	0.6936	0.0506	1	0.0796
F2-5	0.0216	0.0216	0.0988	0.0796	1

F2-1 Appreciation rate of capital preservation
F2-2 Capital accumulation rate
F2-3 Fixed assets growth rate
F2-4 Total assets growth rate
F2-5 Net profit growth rate

Table A.2 The fuzzy similarity matrix of development capability ratios

The Fuzzy Similarity Matrix of Shareholders Profitability Ratios									
	F3-1	F3-2	F3-3	F3-4	F3-5	F3-6	F3-7	F3-8	F3-9
F3-1	1	0.3479	0.2855	0.0953	0.9989	0.0276	0.0593	0.2874	0.0209
F3-2	0.3479	1	0.5293	0.6258	0.3546	0.0689	0.0774	0.1903	0.0226
F3-3	0.2855	0.5293	1	0.0656	0.2867	0.0286	0.0878	0.1708	0.0281
F3-4	0.0953	0.6258	0.0656	1	0.0949	0.0159	0.0172	0.0212	0.0044
F3-5	0.9989	0.3546	0.2867	0.0949	1	0.0277	0.0595	0.2927	0.0209
F3-6	0.0276	0.0689	0.0286	0.0159	0.0277	1	0.4565	0.0445	0.0294
F3-7	0.0593	0.0774	0.0878	0.0172	0.0595	0.4565	1	0.0567	0.0065
F3-8	0.2874	0.1903	0.1708	0.0212	0.2927	0.0445	0.0567	1	0.0234
F3-9	0.0209	0.0226	0.0281	0.0044	0.0209	0.0294	0.0065	0.0234	1

F3-1 Operating income per share
F3-2 Net assets per share
F3-3 Surplus reserve per share
F3-4 Accumulation fund per share
F3-5 Undistributed profit per share
F3-6 Price to book ratio
F3-7 P/E ratio
F3-8 Market to book ratio
F3-9 Return on equity

Table A.3 The fuzzy similarity matrix of shareholders profitability ratios

The Fuzzy Similarity Matrix of Solvency Ratios										
	F4-1	F4-2	F4-3	F4-4	F4-5	F4-6	F4-7	F4-8	F4-9	F4-10
F4-1	1	0.9662	0.0903	0.9161	0.2053	0.1865	0.0028	0.4793	0.0909	0.0902
F4-2	0.9662	1	0.0913	0.8961	0.2181	0.1664	0.0029	0.4582	0.0791	0.0911
F4-3	0.0903	0.0913	1	0.1078	0.1178	0.0322	0.012	0.3261	0.0284	1
F4-4	0.9161	0.8961	0.1078	1	0.2493	0.0924	0.0023	0.4071	0.0673	0.1076
F4-5	0.2053	0.2181	0.1178	0.2493	1	0.0829	0.0111	0.3542	0.3312	0.1176
F4-6	0.1865	0.1664	0.0322	0.0924	0.0829	1	0.0349	0.4995	0.1613	0.0322
F4-7	0.0028	0.0029	0.012	0.0023	0.0111	0.0349	1	0.0773	0.007	0.0119
F4-8	0.4793	0.4582	0.3261	0.4071	0.3542	0.4995	0.0773	1	0.285	0.3259
F4-9	0.0909	0.0791	0.0284	0.0673	0.3312	0.1613	0.007	0.285	1	0.0283
F4-10	0.0902	0.0911	1	0.1076	0.1176	0.0322	0.0119	0.3259	0.0283	1

F4-1 Current ratio
F4-2 Quick ratio
F4-3 Debt ratio
F4-4 Shareholder's equity to liabilities ratio
F4-5 Market value ratio of liabilities to equity
F4-6 fixed assets to total assets ratio
F4-7 Shareholders' equity to total assets ratio
F4-8 Working capital to total assets ratio
F4-9 Working capital to net assets ratio
F4-10 Owner's equity ratio

Table A.4 The fuzzy similarity matrix of solvency ratios

The Fuzzy Similarity Matrix of Operating Capability Ratios									
	F5-1	F5-2	F5-3	F5-4	F5-5	F5-6	F5-7	F5-8	F5-9
F5-1	1	0.0029	0.0053	0.0006	0.0057	0.123	0.0005	0.0033	0.0011
F5-2	0.003	1	0.0353	0.002	0.0171	0.0098	0.0084	0.0171	0.0006
F5-3	0.0053	0.0353	1	0.0073	0.0714	0.0137	0.0703	0.0899	0.0065
F5-4	0.0006	0.002	0.0073	1	0.0067	0.0123	0.0394	0.0356	0.0042
F5-5	0.0057	0.0171	0.0714	0.0067	1	0.0359	0.2368	0.6979	0.0953
F5-6	0.123	0.0098	0.0137	0.0123	0.0359	1	0.432	0.1683	0.0269
F5-7	0.0005	0.0084	0.0703	0.0394	0.2368	0.432	1	0.5784	0.1048
F5-8	0.0033	0.0171	0.0899	0.0356	0.6979	0.1683	0.5784	1	0.1625
F5-9	0.0011	0.0006	0.0065	0.0042	0.0953	0.0269	0.1048	0.1625	1

F5-1 Accounts receivable Turnover
F5-2 Inventory turnover
F5-3 Accounts payable turnover
F5-4 Working capital turnover
F5-5 Current assets turnover
F5-6 Fixed assets turnover
F5-7 Long term assets turnover
F5-8 Total assets turnover
F5-9 Equity turnover

Table A.5 The fuzzy similarity matrix of operating capability ratios

A.2 Construct Fuzzy Equivalence Matrix by Transitive Closure Method

Based on the fuzzy similarity matrix, the fuzzy equivalent matrix is obtained by the transitive closure method. The specific algorithm is based on Subsection 3.6.3 to calculate the profitability, development ability, shareholders' equity, solvency and operating capability fuzzy equivalence matrix.

Show as Table A.6, Table A.7, Table A.8, Table A.9, Table A.10:

The Fuzzy Equivalent Matrix of Profitability Ratios									
	F1-1	F1-2	F1-3	F1-4	F1-5	F1-6	F1-7	F1-8	F1-9
F1-1	1	0.3597	0.3597	0.3597	0.3597	0.2042	0.9134	0.0635	0.0474
F1-2	0.3597	1	0.8196	0.8196	0.8196	0.2042	0.3597	0.06357	0.0474
F1-3	0.3597	0.8196	1	0.9992	0.9904	0.2042	0.3597	0.06357	0.0474
F1-4	0.3597	0.8196	0.9992	1	0.9904	0.2042	0.3597	0.06357	0.0474
F1-5	0.3597	0.8196	0.9904	0.9904	1	0.2042	0.3597	0.06357	0.0474
F1-6	0.2042	0.2042	0.2042	0.2042	0.2042	1	0.2042	0.06357	0.0474
F1-7	0.9134	0.3597	0.3597	0.3597	0.3597	0.2042	1	0.06357	0.0474
F1-8	0.0635	0.0635	0.0635	0.0635	0.0635	0.0635	0.0635	1	0.0474
F1-9	0.0474	0.0474	0.0474	0.0474	0.0474	0.0474	0.0474	0.04749	1

F1-1 Operating gross profit margin
F1-2 Operating income net profit margin
F1-3 Return on assets
F1-4 Net profit rate of total assets
F1-5 Return on current assets
F1-6 Return on fixed assets
F1-7 Profit margin
F1-8 Return on equity
F1-9 Growth rate of main business income

Table A.6 The fuzzy equivalent matrix of profitability ratios

The Fuzzy Equivalent Matrix of Development Capability Ratios					
	F2-1	F2-2	F2-3	F2-4	F2-5
F2-1	1	1	0.0796	0.6936	0.0796
F2-2	1	1	0.0796	0.6936	0.0796
F2-3	0.0796	0.0796	1	0.0796	0.0988
F2-4	0.6936	0.6936	0.0796	1	0.0796
F2-5	0.0796	0.0796	0.0988	0.0796	1

F2-1 Appreciation rate of capital preservation
F2-2 Capital accumulation rate
F2-3 Fixed assets growth rate
F2-4 Total assets growth rate
F2-5 Net profit growth rate

Table A.7 The fuzzy equivalent matrix of development capability ratios

The Fuzzy Equivalent Matrix of Shareholders' Profitability Ratios									
	F3-1	F3-2	F3-3	F3-4	F3-5	F3-6	F3-7	F3-8	F3-9
F3-1	1	0.3546	0.3546	0.3546	0.9989	0.0878	0.0878	0.2927	0.0294
F3-2	0.3546	1	0.5293	0.6258	0.3546	0.0878	0.0878	0.2927	0.0294
F3-3	0.3546	0.5293	1	0.5293	0.3546	0.0878	0.0878	0.2927	0.0294
F3-4	0.3546	0.6258	0.5293	1	0.3546	0.0878	0.0878	0.2927	0.0294
F3-5	0.9989	0.3546	0.3546	0.3546	1	0.0878	0.0878	0.2927	0.0294
F3-6	0.0878	0.0878	0.0878	0.0878	0.0878	1	0.4565	0.0878	0.0294
F3-7	0.0878	0.0878	0.0878	0.0878	0.0878	0.4565	1	0.0878	0.0294
F3-8	0.2927	0.2927	0.2927	0.2927	0.2927	0.0878	0.0878	1	0.0294
F3-9	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	1

F3-1 Operating income per share
F3-2 Net assets per share
F3-3 Surplus reserve per share
F3-4 Accumulation fund per share
F3-5 Undistributed profit per share
F3-6 Price to book ratio
F3-7 P/E ratio
F3-8 Market to book ratio
F3-9 Return on equity

Table A.8 The fuzzy equivalent matrix of shareholders' profitability ratios

The Fuzzy Equivalent Matrix of Solvency Ratios										
	F4-1	F4-2	F4-3	F4-4	F4-5	F4-6	F4-7	F4-8	F4-9	F4-10
F4-1	1	0.9662	0.3261	0.9161	0.3542	0.4793	0.0773	0.4793	0.3312	0.3261
F4-2	0.9662	1	0.3261	0.9161	0.3542	0.4793	0.0773	0.4793	0.3312	0.3261
F4-3	0.3261	0.3261	1	0.3261	0.3261	0.3261	0.0773	0.3261	0.3261	1
F4-4	0.9161	0.9161	0.3261	1	0.3542	0.4793	0.0773	0.4793	0.3312	0.3261
F4-5	0.3542	0.3542	0.3261	0.3542	1	0.3542	0.0773	0.3542	0.3312	0.3261
F4-6	0.4793	0.4793	0.3261	0.4793	0.3542	1	0.0773	0.4995	0.3312	0.3261
F4-7	0.0773	0.0773	0.0773	0.0773	0.0773	0.0773	1	0.0773	0.0773	0.0773
F4-8	0.4793	0.4793	0.3261	0.4793	0.3542	0.4995	0.0773	1	0.3312	0.3261
F4-9	0.3312	0.3312	0.3261	0.3312	0.3312	0.3312	0.0773	0.3312	1	0.3261
F4-10	0.3261	0.3261	1	0.3261	0.3261	0.3261	0.0773	0.3261	0.3261	1

F4-1 Current ratio
F4-2 Quick ratio
F4-3 Debt ratio
F4-4 Shareholder¡¯s equity to liabilities ratio
F4-5 Market value ratio of liabilities to equity
F4-6 fixed assets to total assets ratio
F4-7 Shareholders¡¯ equity to total assets ratio
F4-8 Working capital to total assets ratio
F4-9 Working capital to net assets ratio
F4-10 Owner's equity ratio

Table A.9 The fuzzy equivalent matrix of solvency ratios

The Fuzzy Equivalent Matrix of Operating Capability Ratios									
	F5-1	F5-2	F5-3	F5-4	F5-5	F5-6	F5-7	F5-8	F5-9
F5-1	1	0.0353	0.0899	0.0394	0.123	0.123	0.123	0.123	0.123
F5-2	0.0353	1	0.0353	0.0353	0.0353	0.0353	0.0353	0.0353	0.0353
F5-3	0.0899	0.0353	1	0.0394	0.0899	0.0899	0.0899	0.0899	0.0899
F5-4	0.0394	0.0353	0.0394	1	0.0394	0.0394	0.0394	0.0394	0.0394
F5-5	0.123	0.0353	0.0899	0.0394	1	0.432	0.5784	0.6979	0.1625
F5-6	0.123	0.0353	0.0899	0.0394	0.432	1	0.432	0.432	0.1625
F5-7	0.123	0.0353	0.0899	0.0394	0.5784	0.432	1	0.5784	0.1625
F5-8	0.123	0.0353	0.0899	0.0394	0.6979	0.432	0.5784	1	0.1625
F5-9	0.123	0.0353	0.0899	0.0394	0.1625	0.1625	0.1625	0.1625	1

F5-1 Accounts receivable Turnover
F5-2 Inventory turnover
F5-3 Accounts payable turnover
F5-4 Working capital turnover
F5-5 Current assets turnover
F5-6 Fixed assets turnover
F5-7 Long term assets turnover
F5-8 Total assets turnover
F5-9 Equity turnover

Table A.10 The fuzzy equivalent matrix of operating capability ratios

A.3 Clustering

Assign values from a large to a small to λ , and cluster the fuzzy equivalence matrix. The clustering results of ratios related to profitability, development ability, shareholders' profitability, solvency and Operating capability fuzzy equivalence matrix are shown in Table A.11, Table A.12, Table A.13, Table A.14, Table A.15:

The Clustering result of Profitability Ratios		
λ	Number	Class
1	9	{F1-1}{F1-2}{F1-3}{F1-4}{F1-5}{F1-6}{F1-7}{F1-8}{F1-9}
0.999235	8	{F1-1}{F1-2}{F1-3 F1-4}{F1-5}{F1-6}{F1-7}{F1-8}{F1-9}
0.990468	7	{F1-1}{F1-2}{F1-3}{F1-4} {F1-5}{F1-6}{F1-7}{F1-8}{F1-9}
0.913467	6	{F1-1 F1-7}{F1-2}{F1-3 F1-4 F1-5}{F1-6}{F1-8}{F1-9}
0.819606	5	{F1-1 F1-7}{F1-2 F1-3 F1-4 F1-5}{F1-6}{F1-8}{F1-9}
0.359789	4	{F1-1 F1-2 F1-3 F1-4 F1-5 F1-7}{F1-6}{F1-8}{F1-9}
0.204279	3	{F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7}{F1-8}{F1-9}
0.06357	2	{F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7 F1-8}{F1-9}
0.04749	1	{F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7 F1-8 F1-9}

Table A.11 The clustering result of profitability ratios

The Clustering result of Development Capability Ratios		
λ	Number	Class
1	4	{F2-1 F2-2} {F2-3}{F2-4}{F2-5}
0.693679	3	{F2-1 F2-2 F2-4}{F2-3}{F2-5}
0.09886	2	{F2-1 F2-2 F2-4}{F2-3 F2-5}
0.07966	1	{F2-1 F2-2 F2-3 F2-4 F2-5}

Table A.12 The clustering result of development capability ratios

The Clustering Result of Shareholders' Profitability Ratios		
λ	Number	Class
1	9	{F3-1}{F3-2}{F3-3}{F3-4}{F3-5}{F3-6}{F3-7}{F3-8}{F3-9}
0.998925	8	{F3-1 F3-5}{F3-2}{F3-3}{F3-4}{F3-6}{F3-7}{F3-8}{F3-9}
0.625866	7	{F3-1 F3-5}{F3-2 F3-4}{F3-3}{F3-6}{F3-7}{F3-8}{F3-9}
0.529356	6	{F3-1 F3-5}{F3-2 F3-3 F3-4}{F3-6}{F3-7}{F3-8}{F3-9}
0.456587	5	{F3-1 F3-5}{F3-2 F3-3 F3-4}{F3-6 F3-7}{F3-8}{F3-9}
0.354634	4	{F3-1 F3-2 F3-3 F3-4 F3-5}{F3-6 F3-7}{F3-8}{F3-9}
0.292766	3	{F3-1 F3-2 F3-3 F3-4 F3-5 F3-8}{F3-6 F3-7}{F3-9}
0.087876	2	{F3-1 F3-2 F3-3 F3-4 F3-5 F3-6 F3-7 F3-8}{F3-9}
0.029443	1	{F3-1 F3-2 F3-3 F3-4 F3-5 F3-6 F3-7 F3-8 F3-9}

Table A.13 The clustering result of shareholders' profitability ratios

The Clustering Result of Solvency Ratios		
λ	Number	Class
1	10	{F4-1}{F4-2}{F4-3}{F4-4}{F4-5}{F4-6}{F4-7}{F4-8}{F4-9}{F4-10}
0.999994	9	{F4-1}{F4-2}{F4-3 F4-10}{F4-4}{F4-5}{F4-6}{F4-7}{F4-8}{F4-9}
0.966226	8	{F4-1 F4-2}{F4-3 F4-10}{F4-4}{F4-5}{F4-6}{F4-7}{F4-8}{F4-9}
0.916107	7	{F4-1 F4-2 F4-4}{F4-3 F4-10}{F4-5}{F4-6}{F4-7}{F4-8}{F4-9}
0.499507	6	{F4-1 F4-2 F4-4}{F4-3 F4-10}{F4-5}{F4-6 F4-8}{F4-7}{F4-9}
0.479341	5	{F4-1 F4-2 F4-4 F4-6 F4-8}{F4-3 F4-10}{F4-5}{F4-7}{F4-9}
0.354197	4	{F4-1 F4-2 F4-4 F4-5 F4-6 F4-8}{F4-3 F4-10}{F4-7}{F4-9}
0.331237	3	{F4-1 F4-2 F4-4 F4-5 F4-6 F4-8 F4-9}{F4-3 F4-10}{F4-7}
0.326149	2	{F4-1 F4-2 F4-3 F4-4 F4-5 F4-6 F4-8 F4-9 F4-10}{F4-7}
0.077268	1	{F4-1 F4-2 F4-3 F4-4 F4-5 F4-6 F4-7 F4-8 F4-9 F4-10}

Table A.14 The clustering result of solvency ratios

The clustering result of Operating Capability Ratios		
λ	Number	Class
1	9	{F5-1}{F5-2}{F5-3}{F5-4}{F5-5}{F5-6}{F5-7}{F5-8}{F5-9}
0.697926	8	{F5-1}{F5-2}{F5-3}{F5-4}{F5-5 F5-8}{F5-6}{F5-7}{F5-9}
0.578484	7	{F5-1}{F5-2}{F5-3}{F5-4}{F5-5 F5-7 F5-8}{F5-6}{F5-9}
0.432017	6	{F5-1}{F5-2}{F5-3}{F5-4}{F5-5 F5-6 F5-7 F5-8}{F5-9}
0.162565	5	{F5-1}{F5-2}{F5-3}{F5-4}{F5-5 F5-6 F5-7 F5-8 F5-9}
0.123092	4	{F5-1 F5-5 F5-6 F5-7 F5-8 F5-9}}{F5-2}{F5-3}{F5-4}
0.089924	3	{F5-1 F5-3 F5-5 F5-6 F5-7 F5-8 F5-9}{F5-2}{F5-4}
0.039425	2	{F5-1 F5-2 F5-3 F5-5 F5-6 F5-7 F5-8 F5-9}{F5-4}
0.035325	1	{F5-1 F5-2 F5-3 F5-4 F5-5 F5-6 F5-7 F5-8 F5-9}

Table A.15 The clustering result of operating capability ratios

A.4 Screen the Ratios in Same Classification based on Correlation Coefficient

We will choose 3 ratios by using correlation coefficient from profitability ratios, development capability ratios, shareholders profitability ratios, solvency ratios and operating capability ratios. As a result, we will have 15 ratios in total.

A.4.1 Profitability ratios screening

Choose the classification when λ=0.204279 as a start point, so the classification is F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7, F1-8, F1-9, F1-8 F1-9 are already single ratios in single classes, so both can be counted as one of the final 15 ratios. However, F1-1 F1-2 F1-3 F1-4 F1-5 F1-6 F1-7 are in the same class. We need to choose one ratio that can represent the class by using the correlation coefficient method.

A.4.2 Calculate correlation coefficient A_{ij}

Correlation Coefficients 1							
A_{ij}	F1-1	F1-2	F1-3	F1-4	F1-5	F1-6	F1-7
F1-1	1	-0.13	0.0338	0.0377	0.0236	0.0161	-0.9135
F1-2	-0.13	1	0.8159	0.818	0.8196	0.2043	0.3598
F1-3	0.0338	0.8159	1	0.9992	0.9905	0.1566	0.1586
F1-4	0.0377	0.818	0.9992	1	0.9894	0.1587	0.1586
F1-5	0.0236	0.8196	0.9905	0.9894	1	0.1453	0.1619
F1-6	0.0161	0.2043	0.1566	0.1587	0.1453	1	0.0278
F1-7	-0.9135	0.3598	0.1586	0.1586	0.1619	0.0278	1

F1-1 Operating gross profit margin
F1-2 Operating income net profit margin
F1-3 Return on assets
F1-4 Net profit rate of total assets
F1-5 Return on current assets
F1-6 Return on fixed assets
F1-7 Profit margin

Table A.16 Correlation coefficients 1

A.4.3 Calculate correlation index R

$$
\begin{aligned}
R_{A1} &= a_{1,2}^2 + a_{1,3}^2 + a_{1,4}^2 + a_{1,5}^2 + a_{1,6}^2 + a_{1,7}^2 = 0.1424 \\
R_{A2} &= a_{2,1}^2 + a_{2,3}^2 + a_{2,4}^2 + a_{2,5}^2 + a_{2,6}^2 + a_{2,7}^2 = 0.3658 \\
R_{A3} &= a_{3,1}^2 + a_{3,2}^2 + a_{3,4}^2 + a_{3,5}^2 + a_{3,6}^2 + a_{3,7}^2 = 0.4493 \\
R_{A4} &= a_{4,1}^2 + a_{4,2}^2 + a_{4,3}^2 + a_{4,5}^2 + a_{4,6}^2 + a_{4,7}^2 = 0.4497 \\
R_{A5} &= a_{5,1}^2 + a_{5,2}^2 + a_{5,3}^2 + a_{5,4}^2 + a_{5,6}^2 + a_{5,7}^2 = 0.4466 \\
R_{A6} &= a_{6,1}^2 + a_{6,2}^2 + a_{6,3}^2 + a_{6,4}^2 + a_{6,5}^2 + a_{6,7}^2 = 0.0189 \\
R_{A7} &= a_{7,1}^2 + a_{7,2}^2 + a_{7,3}^2 + a_{7,4}^2 + a_{7,5}^2 + a_{7,6}^2 = 0.1735
\end{aligned}
$$

A.4.4 Screening

From the result we can conclude that R_{A4} is the largest correlation index, so F1-4 is picked into the final feature sets.

After screening, the 3 ratios in the profitability category are: F1-8 Return on equity, F1-9 Growth rate of main business income, F1-4 Net profit rate of total assets.

1. Development capability ratios screening

   Choose the classification when λ=0.331237, so the classification is F2-1 F2-2 F2-4F2-3F2-5, F2-3, F2-5 are already a single ratios in the single class so both can be counted as one of the final 15 ratios. However, F2-1 F2-2 F2-4 are in the same class, we need to choose one ratio that can represent the class by using correlation coefficient method.

   (a) Calculate correlation coefficient B_{ij}

Correlation Coefficients 2			
B_ij	F2-1	F2-2	F2-4
F2-1	1	1	0.6937
F2-2	1	1	0.6937
F2-4	0.6937	0.6937	1
F2-1 Appreciation rate of capital preservation			
F2-2 Capital accumulation rate			
F2-4 Total assets growth rate			

Table A.17 Correlation coefficients 2

   (b) Calculate correlation index R

$$
\begin{aligned}
R_{B1} &= b_{1,2}^2 + b_{1,4}2 = 0.7406 \\
R_{B2} &= b_{2,1}^2 + b_{2,4}^2 = 0.7406 \\
R_{B4} &= b_{4,1}^2 + b_{4,2}^2 = 0.4812
\end{aligned}
$$

(c) Screening

From the result we can conclude that R_{B1} or R_{B2} is the largest correlation index, so we can pick any one of F2-1 and F2-2 into the final feature sets, here we pick F2-1.

After screening, the 3 ratios in development capability category are: F2-3 Fixed assets growth rate, F2-5 Net profit growth rate, F2-1 Appreciation rate of capital preservation.

2. Share holders' equity ratios screening

Choose the classification when λ=0.292766, so the classification is F3-1 F3-2 F3-3 F3-4 F3-5 F3-8F3-6 F3-7, F3-9 is already a single ratio in a single class so both can be counted as one of the final 15 ratios. However, F3-1 F3-2 F3-3 F3-4 F3-5 F3-8 is in the same class, so as F3-6 F3-7. we need to choose one ratio from each of the class that can represent the class by using correlation coefficient method.

Correlation Coefficients 3						
C_ij	F3-1	F3-2	F3-3	F3-4	F3-5	F3-8
F3-1	1	0.3479	0.2855	0.0954	0.9989	0.2874
F3-2	0.3479	1	0.5294	0.6259	0.3546	0.1903
F3-3	0.2855	0.5294	1	0.0656	0.2867	0.1708
F3-4	0.0954	0.6259	0.0656	1	0.0949	0.0212
F3-5	0.9989	0.3546	0.2867	0.0949	1	0.2927
F3-8	0.2874	0.1903	0.1708	0.0212	0.2927	1
F3-1 Operating income per share						
F3-2 Net assets per share						
F3-3 Surplus reserve per share						
F3-4 Accumulation fund per share						
F3-5 Undistributed profit per share						
F3-8 Market to book ratio						

Table A.18 Correlation coefficients 3

$$R_{C1} = c_{1,2}^2 + c_{1,3}^2 + c_{1,4}^2 + c_{1,5}^2 + c_{1,8}^2 = 0.2584$$
$$R_{C2} = c_{2,1}^2 + c_{2,3}^2 + c_{2,4}^2 + c_{2,5}^2 + a_{2,8}^2 = 0.1910$$
$$R_{C3} = c_{3,1}^2 + c_{3,2}^2 + c_{3,4}^2 + c_{3,5}^2 + c_{3,8}^2 = 0.0955$$
$$R_{C4} = c_{4,1}^2 + c_{4,2}^2 + c_{4,3}^2 + c_{4,5}^2 + c_{4,8}^2 = 0.0829$$
$$R_{C5} = c_{5,1}^2 + c_{5,2}^2 + c_{5,3}^2 + c_{5,4}^2 + c_{5,8}^2 = 0.2601$$
$$R_{C8} = c_{8,1}^2 + c_{8,2}^2 + c_{8,3}^2 + c_{8,4}^2 + c_{8,6}^2 = 0.0468$$

From the result we can see that F3-5 should be chosen.

Correlation Coefficients 4		
C_{ij}	F3-6	F3-7
F3-6	1	0.4566
F3-7	0.4566	1
F3-6 Price to book ratio		
F3-7 P/E ratio		

Table A.19 Correlation coefficients 4

$$R_{A6} = c_{6,7}^2 = 0.2085$$
$$R_{A7} = a_{7,6}^2 = 0.2085$$

The correlation index of two ratios are equivalent, so here we can choose F3-6. After screening, the 3 ratios in shareholders' profitability category are: F3-5 Undistributed profit per share, F3-6 Price to book ratio and F3-9 Return on equity.

3. Solvency ratios screening

   Choose the classification when λ=0.292766, so the classification is F4-1 F4-2 F4-4 F4-5 F4-6 F4-8 F4-9F4-3 F4-10F4-7, F4-7 is already a single ratio in a single class so it can be count as one of the final 15 ratios. However, F4-1 F4-2 F4-4 F4-5 F4-6 F4-8 F4-9 are in the same class, so as F4-3 F4-10. we need to choose one ratio from each of the class that can represent the class by using correlation coefficient method. The steps is just same as we did previously.

$$R_{D1} = d_{1,2}^2 + d_{1,4}^2 + d_{1,5}^2 + d_{1,6}^2 + d_{1,8}^2 + d_{1,9}^2 = 0.3480$$
$$R_{D2} = d_{2,1}^2 + d_{2,4}^2 + d_{2,5}^2 + d_{2,6}^2 + d_{2,8}^2 + d_{2,9}^2 = 0.3380$$
$$R_{D4} = d_{4,1}^2 + d_{4,2}^2 + d_{4,5}^2 + d_{4,6}^2 + d_{4,8}^2 + d_{4,9}^2 = 0.3139$$
$$R_{D5} = d_{5,1}^2 + d_{5,2}^2 + d_{5,4}^2 + d_{5,6}^2 + d_{5,8}^2 + d_{5,9}^2 = 0.0657$$
$$R_{D6} = d_{6,1}^2 + d_{6,2}^2 + d_{6,4}^2 + d_{6,5}^2 + d_{6,8}^2 + d_{6,9}^2 = 0.0589$$
$$R_{D8} = d_{8,1}^2 + d_{8,2}^2 + d_{8,4}^2 + d_{8,5}^2 + d_{8,6}^2 + d_{8,9}^2 = 0.1769$$
$$R_{D9} = d_{9,1}^2 + d_{9,2}^2 + d_{9,4}^2 + d_{9,5}^2 + d_{9,6}^2 + d_{9,8}^2 = 0.0393$$

From the result we can see that F4-1 should be chosen. As a result the 3 ratios in shareholders' profitability category should be: F4-3 Debt ratio, F4-7 Shareholders' equity to total assets ratio, F4-1 Current ratio.

Correlation Coefficients 5							
D_{ij}	F4-1	F4-2	F4-4	F4-5	F4-6	F4-8	F4-9
F4-1	1	0.9662	0.9161	-0.2053	-0.1865	0.4793	0.0908
F4-2	0.9662	1	0.8961	-0.2181	-0.1664	0.4582	0.0791
F4-4	0.9161	0.8961	1	-0.2493	-0.0923	0.4071	0.0673
F4-5	-0.2053	-0.2181	-0.2493	1	0.0828	-0.3542	-0.3312
F4-6	-0.1865	-0.1664	-0.0923	0.0828	1	-0.4995	-0.1613
F4-8	0.4793	0.4582	0.4071	-0.3542	-0.4995	1	0.285
F4-9	0.0908	0.0791	0.0673	-0.3312	-0.1613	0.285	1

F4-1 Current ratio
F4-2 Quick ratio
F4-4 Shareholder's equity to liabilities ratio
F4-5 Market value ratio of liabilities to equity
F4-6 fixed assets to total assets ratio
F4-8 Working capital to total assets ratio
F4-9 Working capital to net assets ratio

Table A.20 Correlation coefficients 5

4. Operating capability ratios screening

   Choose the classification when λ=0.089924, the classification in this value is F5-1 F5-3 F5-5 F5-6 F5-7 F5-8 F5-9F5-2F5-4, F5-2 and F5-4 are already in a single class so they can be count as one of the final 15 ratios directly. However, F5-1 F5-3 F5-5 F5-6 F5-7 F5-8 F5-9 is in the same class. We need to choose one ratio from the class that can represent the class by using correlation coefficient method. The steps is same as we did previously.

$$
\begin{aligned}
R_{E1} &= e_{1,3}^2 + e_{1,5}^2 + e_{1,6}^2 + e_{1,7}^2 + e_{1,8}^2 + e_{1,9}^2 = 0.0025 \\
R_{E3} &= e_{3,1}^2 + e_{3,5}^2 + e_{3,6}^2 + e_{3,7}^2 + e_{3,8}^2 + e_{3,9}^2 = 0.0031 \\
R_{E5} &= e_{5,1}^2 + e_{5,3}^2 + e_{5,6}^2 + e_{5,7}^2 + e_{5,8}^2 + e_{5,9}^2 = 0.0931 \\
R_{E6} &= e_{6,1}^2 + e_{6,3}^2 + e_{6,5}^2 + e_{6,7}^2 + e_{6,8}^2 + e_{6,9}^2 = 0.0387 \\
R_{E7} &= e_{7,1}^2 + e_{7,3}^2 + e_{7,5}^2 + e_{7,6}^2 + e_{7,8}^2 + e_{7,9}^2 = 0.0989 \\
R_{E8} &= e_{8,1}^2 + e_{8,3}^2 + e_{8,5}^2 + e_{8,6}^2 + e_{8,7}^2 + e_{8,9}^2 = 0.1474 \\
R_{E9} &= e_{9,1}^2 + e_{9,3}^2 + e_{9,5}^2 + e_{9,6}^2 + e_{9,7}^2 + e_{9,8}^2 = 0.0079
\end{aligned}
$$

   From the result we can see that F5-8 should be chosen. As a result the 3 ratios in operating capability catagory should be: F5-2 Inventory turnover F5-4 Working capital turnover F5-8 Total assets turnover.

Correlation Coefficients 6							
E_{ij}	F5-1	F5-3	F5-5	F5-6	F5-7	F5-8	F5-9
F5-1	1	0.0054	0.0057	0.1231	0.0005	0.0033	-0.0011
F5-3	0.0054	1	0.0714	0.0138	0.0703	0.0899	0.0065
F5-5	0.0057	0.0714	1	0.036	0.2368	0.6979	0.0953
F5-6	0.1231	0.0138	0.036	1	0.432	0.1683	0.0269
F5-7	0.0006	0.0703	0.2368	0.432	1	0.5784	0.1048
F5-8	0.0033	0.0899	0.6979	0.1683	0.5784	1	0.1625
F5-9	-0.0011	0.0065	0.0953	0.0269	0.1048	0.1625	1

F5-1 Accounts receivable Turnover
F5-3 Accounts payable turnover
F5-5 Current assets turnover
F5-6 Fixed assets turnover
F5-7 Long term assets turnover
F5-8 Total assets turnover
F5-9 Equity turnover

Appendix B

Data Platform Access

B.1 Financial Ratios Access

The financial ratios are calculated from the listed company's annual report downloaded from the China Stock Market Accounting Research Database (CSMAR).

1. Log in to the CSMAR website, http://www.gtarsc.com
2. Enter the 'Chinese Listed Companies Financial Index Analysis Database.'
3. Most of the financial indicators used in this research can be extracted from the database, and some of the indicators need to be calculated separately based on the relation between the indicators and the indicator formula.

B.2 Stock price, Volume and Technical ratios access

1. Download and log in to Tongdaxin stock trading platform, https://www.tdx.com.cn/
2. Select the data download after market close in 'Settings', and download the required stock historical data
3. Data export
4. Calculate technical indicators based on exported data

Appendix C

Parameters Table

Table of Parameters Setting	
Random Forest	
Tree_numbers	100
Node size	5
Sample size	0.632 of the total size
Genetic Algorithm	
Population size	100
Terminated evolutionary generation	20
Crossover probability	0.7
Mutation probability	Pm = 0.7/Lind, where Lind is the chromosome length
Support vector regression	
Search space of C	0.1-10
Search space of g	0.1-10
Search space of p	0.01-1