Journal of Investment and Management
2018; 7(4): 117-124
http://www.sciencepublishinggroup.com/j/jim
doi: 10.11648/j.jim.20180704.12
ISSN: 2328-7713 (Print); ISSN: 2328-7721 (Online)

A New Stock Selection Model Based on Decision Tree C5.0 Algorithm

Qiansheng Zhang, Jingru Zhang, Zisheng Chen, Miao Zhang, Songying Li

School of Finance, Guangdong University of Foreign Studies, Guangzhou, P.R. China

Email address:
zhqiansh01@126.com (Qiansheng Zhang)

To cite this article:
Qiansheng Zhang, Jingru Zhang, Zisheng Chen, Miao Zhang, Songying Li. A New Stock Selection Model Based on Decision Tree C5.0 Algorithm. *Journal of Investment and Management*. Vol. 7, No. 4, 2018, pp. 117-124. doi: 10.11648/j.jim.20180704.12

Received: August 10, 2018; **Accepted**: September 1, 2018; **Published**: September 21, 2018

Abstract: Due to the disordered characteristic and strong randomness of China's stock market, the typical data mining algorithms currently used to analyze and forecast the stock have imprecise prediction outcomes. In order to solve this problem, based on the industry rotation cycle theory, this paper constructs a new stock selection model combining Decision Tree C5.0 Algorithm and factor analysis. Industry rotation cycle theory aims to analyze the development trend of various industries to find promising industries as initial stock pool. According to this principle, this paper selects four industries and the A-share stocks of these industries are used as initial stock pool. This paper builds a stock index system consisting of six effective factors based on the factor analysis of stocks financial indicators and technical indicators. Then Decision Tree C5.0 Algorithm is presented to realize the prediction of stock returns and the classification of stocks. The empirical test of the proposed stock selection model, using the data from the second and the third quarter of 2017 in China A-share stock market, demonstrates that this model has significant difference in the classification accuracy between low-yielding stocks and high-yielding stocks in that case classification accuracy shows a trend opposite against stock return rate. In a conclusion, this model can effectively help investors to avoid risks and make rational investment but has little effect on obtaining excess return.

Keywords: Decision Tree C5.0, Factor Analysis, Stock Selection Model Introduction

1. Introduction

With the development of information technology, quantitative analysis has gradually become the mainstream investment analysis method. In China's unique investment market environment, there is still huge exploration space and application potential for quantitative analysis practice. How to use quantitative models to select portfolios which can exceed benchmark returns and maintain stable volatility has become the focus of many institutional and individual investors.

There are many kinds of quantitative analysis tools, among which decision tree has advantages such as strong anti-noise capability, Good contractility and clear classification form, so it is widely used in financial investment decision-making. Wang Dong et al [1] used rough set theory on each split attribute of decision tree to improve the selection accuracy, solving the disadvantages of local optimal solution and effectively improving the prediction accuracy. Dharini et al [2] combined the decision tree with the fuzzy system. After using the decision tree to extract eigenvalue, the adaptive neural fuzzy system was employed to stock classification. The experiment results showed that the new algorithm was more accurate than the traditional decision tree algorithm. Mantri [3] used decision tree and support vector machine(SVM) to analyze and forecast Mumbai index based on historical data of Bombay stock exchange respectively. The results revealed that decision tree was superior to SVM in data verification. Sasbarakan [4] combined several classification models together to obtain more accurate prediction of stock price and risk. Experiments demonstrated that the combination of Bagging model and decision tree can effectively increase prediction accuracy. The research on decision-making tree started late in China, but a lot of achievement have been accomplished. Wei [5] established a binary decision tree to predict the short-term trend of stocks, helping investors choose stocks. Huang [6] developed the ID3 decision tree algorithm and then used this model to classify and predict stock, proving that it was feasible and effective to use ID3 algorithm to realize classification and prediction of stock in China's market. Zhang [7] selected Shanghai and Shenzhen securities as the initial stock pool, and then screened the

indicators related to the stock trend, finally classified the stocks by the decision tree and concluded a stock selection strategy. Tao [8] selected 200 listed companies as samples and 14 representative indicators as input variables and respectively established three classification models including C5.0 decision tree, BP neural network and RBF neural network. The results proved that the classification prediction effect of C5.0 was optimal. Yang [9] optimized the stock technical indexes system and C4.5 decision tree. It showed that the optimized C4.5 algorithm could help investors choose stocks with higher returns. Huang [10] searched for stocks with high investment value with mixed method based on association rules algorithm in data mining, decision tree model and neural network model. The results showed that both decision tree and neural network performed well in stock selection. Shen [11] established CART decision tree stock selection model and classification model of financial indexes of listed companies to analyze and forecast of the yield rate of stock.

Because the decision tree has the advantages of strong Generalization, this paper establishes the C5.0 decision tree multi-factor stock selection model. On the basis of analyzing and studying the domestic and abroad stock indexes system, this paper selects technical indexes and financial indexes respectively to construct stock index system. In order to reduce the volatility of stock returns, this paper first judges the macroeconomic trend and analyzes the industry cycle based on the industry rotation theory. China A-shares stocks in electronic industry, machinery industry and other industries are selected as the initial stock pool. Finally, the quarterly data from 2016 to 2017 are used to train and test the stock selection model. Empirical results show that the stock selection model has high classification accuracy and Good adaptability to China's stock market. Therefore, this paper provides a simple and convenient investment strategy with low risk for investors.

2. Initial Stock Pool

The industry choice is crucial to return rate of the investment portfolio. The source of profit of the portfolio is the intrinsic value of industry which is closely related to the macro-economic cycle. This paper adopts industry rotation theory to select industries with Good development trend which are classified by the standard of China Securities Regulatory Commission (CSRC).

Every industry has its own development rules. The development process of an industry from Generation to prosperity and then to wane is called industry lifecycle [12]. Under the influence of the macro-economic cycle, the industry can be classified according to the industry lifecycle and an index system can be established to analyze the stage of industry lifecycle and prospect of the industry [13]. At present, China is in the stage of economic recovery. The industry structure and consumption structure are Gradually transforming, and the performance of periodic industries is more prominent than other industries. At the same time, China's demographic dividend is about to disappear, accompanying with the decline of human resources in manufacturing industry, and the steady Growth of capital intensive industries targeting on technology innovation. According to the industry cycle theory, this paper analyzes the following industries.

The pharmaceutical industry belongs to the young periodic cycle industry which has always been regarded as a sunrise industry [14]. Since China is currently in the stage of industry transformation and consumption transformation, the pharmaceutical industry, as an industry has little correlation with the macro-environment, deserves investors' attention. At the same time, due to the stimulation of a series of social phenomena and policies such as aging and the two-child policy, the pharmaceutical industry is expected to have further profit Growth space.

As industries producing diversified merchandises, the electronic and information technology industry has a promising future. With the rapid expansion of new artificial intelligent algorithm such as big data and cloud computing, the information technology industry has gradually played a leading role in the national economy. Meanwhile, a large amount of information technology has been widely used in the electronic industry, further promoting the development of the electronic industry. From the perspective of the industry lifecycle, the information technology industry is at the stage of emerging and developing vigorously. The electronic industry is developing steadily for many years and the profit margin is increased stably. Therefore, the investment in the electronic and information technology industry is worthy of attention.

Machinery industry, as the mainstay of China's industry, maintains important strategic position whose market demand increases steadily and prospect is promising. Although the overall situation of machinery industry closely links to the China's macro economy environment, this industry has always been controlled and regulated by the Government, small fluctuations exist, but the general developing trend is favorable.

Therefore, according to the above analysis, this paper selects 1548 stocks in China A-share stock market of pharmaceutical industry, electronics industry, machinery industry, and information technology industry as the initial stock pool.

3. Establishment of Stock Index System

3.1. Index Selection

Since there are many factors fluctuating stock price, the diversity of factors and the economic significance reflected by each factor should be taken into full account when selecting candidate indexes.

Growth indicator is measurement of a company's future development, that is, the potential of the company. In this paper, representative Growth indicators include total assets Growth rate, earnings per share Growth rate, net margin Growth rate and revenue Growth rate.

Cash flow indicator measures a company's ability to sustain operations over a long period of time. The expansion and operation of a company is related to the cash flow. If a company's cash flow is difficult to turn over, it will has a Great

impact on the daily operation of the company, even causes the company to shut down or Go bankrupt. On the contrary, stable and sufficient cash flow will help the company expand its business scope and reduce risks, thus improving the intrinsic value of the company. In this paper, cash flow ratio and net cash flow Growth rate from operating activities are included in cash flow indicators.

In order to reflect the influence of market sentiment on investors' behavior, this paper introduces some technical indicators which directly reflect the market cycle and indirectly reflect the behavior and sentiment of investors, thus forecasting the future changes of the market. In this paper, the shareholder change rate is selected as the factor of technical analysis.

The operating indicators measure the operating efficiency of a company, that is, the efficiency and benefit Gained from operation of assets and capital. This paper selects rate of return on total assets and total assets turnover rate.

Table 1. *Stock index system.*

Growth indicators	Cash flow indicators	Technical indicators	Operating indicators
Total Asset Growth Rate	Cash Flow Ratio	Shareholder Change Rate	Rate of Return on Total Assets
Earnings per Share Growth Rate	Net Cash Flow Growth Rate from Operating Activities		Total Assets Turnover Rate
Net Margin Growth Rate			
Revenue Growth Rate			

Then the stocks in the initial stock pool should be filtered to prepare for the following analysis. Firstly, seeing that the particularity of stock indicators and yield rate of new stocks, stocks which listed for less than one year are supposed to be excluded; secondly, the stocks which are suspended or cannot be traded in market day are excluded. Financial statements are generally published four times a year, that is, first quarterly report, semiannual report, third quarterly report and annual report. The figures for the second and fourth quarters can be calculated based on the data of the four quarters. Therefore, this paper chooses one quarter as a unit to analyze.

3.2. Factor Analysis

The correlation between the single factor and the yield rate of the next quarterly period is tested by sorting method. Factors which can pass the empirical test are selected from the initial index system and add into waiting list of the next test.

In the first step, single stock index of each stock in each time period is sorted separately according to the value of each stock index and divided into five Groups. Then the average excess return rate of each Group of stocks is calculated. The following are the results of all candidate factors, where the horizontal coordinate represents the candidate factor index and vertical coordinate represents return rate.

AER is the abbreviation of average excess return.

Growth Indicators

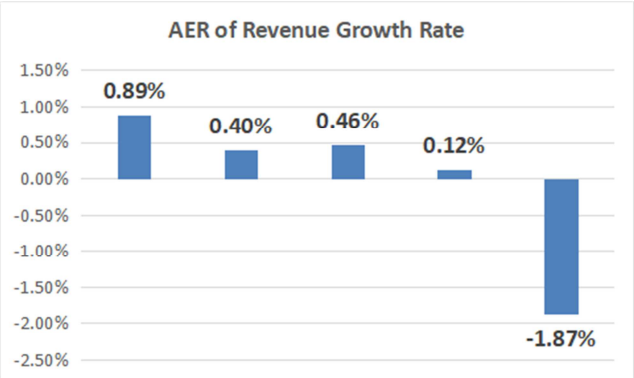

Figure 1. *AER of revenue growth rate.*

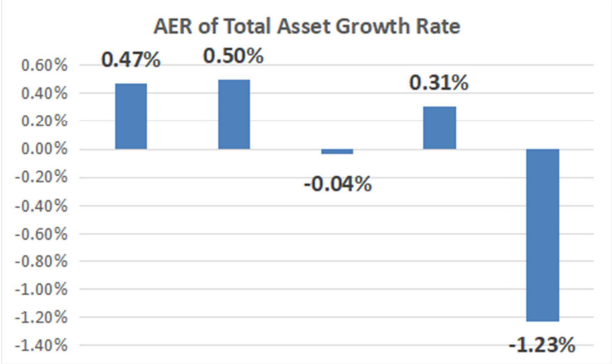

Figure 2. *AER of total asset growth rate.*

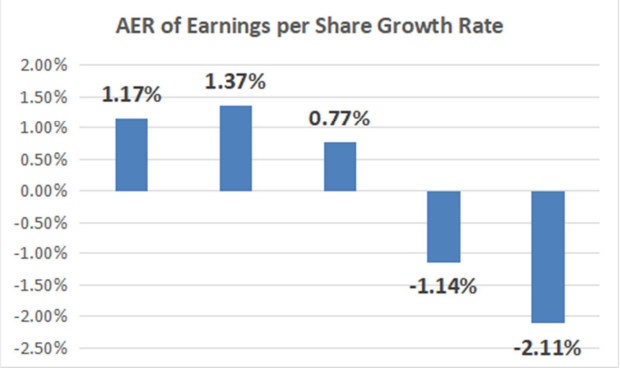

Figure 3. *AER of earnings per share growth rate.*

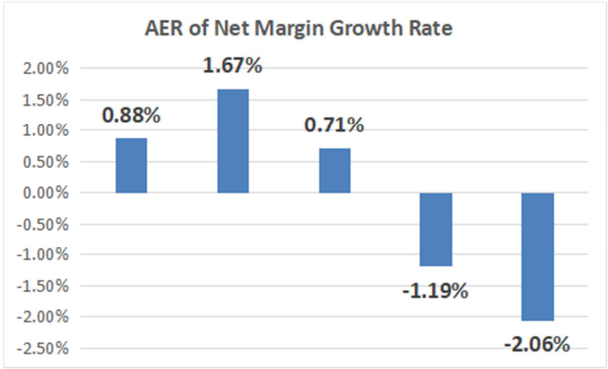

Figure 4. *AER of net margin growth rate.*

Cash Flow Indicators

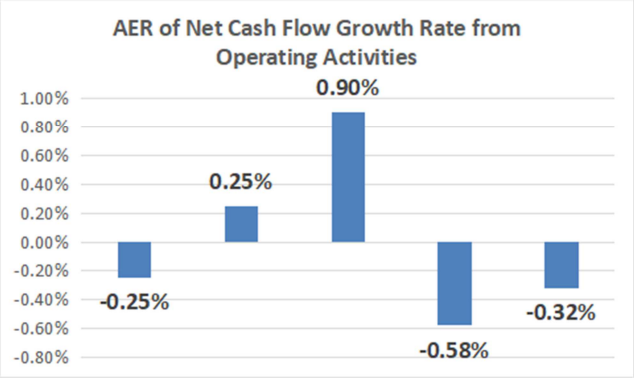

Figure 5. *AER of Net Cash Flow Growth Rate from Operating Activities.*

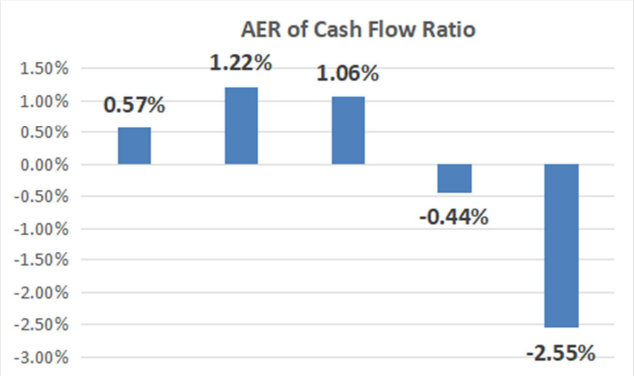

Figure 6. *AER of Cash Flow Ratio.*

Technological Indicators

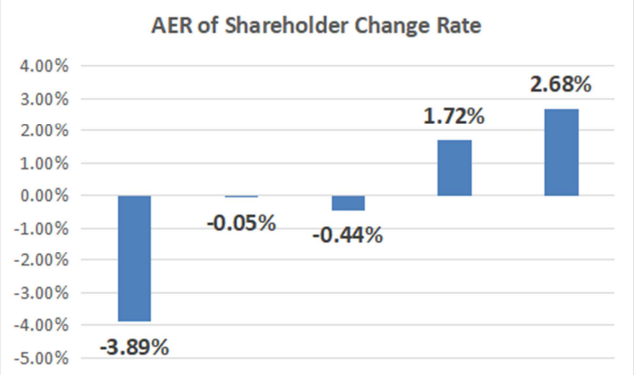

Figure 7. *AER of Shareholder Change Rate.*

Operating Indicators

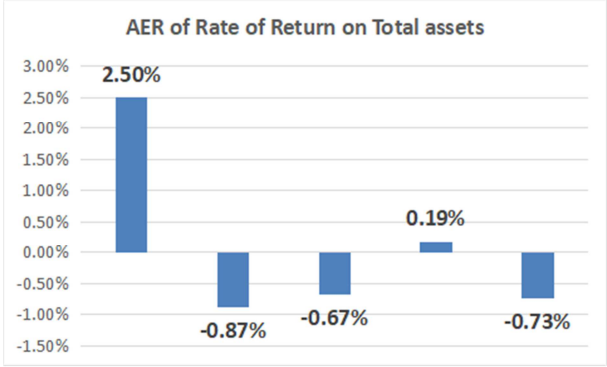

Figure 8. *AER of rate of return on total assets.*

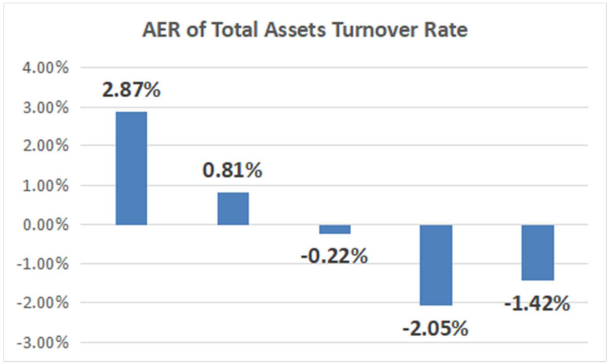

Figure 9. *AER of total assets turnover rate.*

Secondly, the correlation between the stock indexes and the average excess return rate of each Group is examined. The results show that all the stock indexes are tested, except the net cash flow Growth rate from operating activities and the rate of return on total assets whose correlation test results are less than 0.75.

Table 2. *The Results of Correlation Test.*

Factors	Correlation Coefficient
Total Assets Growth Rate	0.787
Earnings Per Share Growth Rate	0.926
Net Margin Growth Rate	0.886
Revenue Growth Rate	0.847
Cash Flow Rate	0.807
Net Cash Flow Growth Rate from operating Activities	0.262
Shareholder Change Rate	-0.934
Rate of Return on Total assets	0.605
Total Assets Turnover Rate	0.929

Table 3. *Results of significance test.*

Factor	T Value	Significance
Total Assets Growth Rate	1.3021	Not Significant
Earnings Per Share Growth Rate	3.3530	Significant on the level of 5%
Net Margin Growth Rate	2.3363	Significant on the level of 10%
Revenue Growth Rate	2.3547	Significant on the level of 10%
Cash Flow Rate	2.2204	Significant on the level of 10%
Shareholder Change Rate	-3.5177	Significant on the level of 5%
Total Assets Turnover Rate	3.1705	Significant on the level of 5%

Finally, after observing the frequency that the yield rate of the fifth Group is higher than that of the first Group, the average excess return rate of the low-ranking Group is subtracted from that factor. Conduct one sample t test on the sequence and α is 0.1. If the sequence can pass the significance test, it means that a significant difference exists between the yield return of the low-ranking Group and the high-ranking Group. Except for the total assets Growth rate, all the other six factors pass the significance test.

3.3 Redundancy Test of Factors

Due to the inherent consistency of each factor, a relatively high correlation maybe occurs between some indexes. To avoid this problem, it is necessary to retain the most significant index among relevant indexes, and the rest of indexes need to be eliminated as redundant indexes. The Spearman rank correlation coefficient is used for redundancy analysis:

$$\rho = 1 - \frac{6\sum d_i^2}{n^3 - n}$$

Table 4. *Results of redundancy test.*

	Total Assets Turnover Rater	Revenue Growth Rate	Cash Flow Rate	Net Margin Growth Rate	Earnings Per Share Growth Rate	Shareholder Change Rate
Total Assets Turnover Rate	1.00	0.04	0.11	0.03	0.03	-0.02
Revenue Growth Rate	0.04	1.00	-0.02	0.51	0.46	-0.01
Cash Flow Rate	0.11	-0.02	1.00	0.01	0.01	-0.03
Net Margin Growth Rate	0.03	0.51	0.01	1.00	0.88	-0.03
Earnings Per Share Growth Rate	0.03	0.46	0.01	0.88	1.00	-0.07
Shareholder Change Rate	-0.02	-0.01	-0.03	-0.03	-0.07	1.00

According to the results, the Spearman rank correlation coefficient between the net margin Growth rate and the earnings per share Growth rate reaches 0.88, indicating that these two factors are mutually substituted indexes and only one needs be remained. In order to show the impact of the quarterly period on the stock, the quarter is added into the stock index system. To sum up, the stock selection indexes are: Total Assets Turnover Rate, Revenue Growth Rate, Cash Flow Rate, Earnings Per Share Growth Rate, Shareholder Change Rate, and the quarter.

4. Decision Tree C5.0 Algorithm

Decision tree is a common classification method which is based on the probability and statistics by establishing rules to classify samples using recursive program from top to bottom. The nodes of decision tree can be divided into end node and chance node as well as decision node. Chance node represents an attribute or a feature whose different values or figures lead to different categories. Therefore, the decision tree tests the individual attribute values of different nodes until it reaches the final category of each sample, which is the end node. Through this process, each path from decision node to end node can be regarded a rule and multiple rules construct an entire decision tree.

Quinlan proposed ID3 algorithm [15] which mainly aimed to classify discrete data in 1973. However, ID3 algorithm selects and evaluates attributes based on Kullback-Leibler divergence, in other words, relative entropy or information divergence, resulting in neglecting the splitting attributes with less values. In 1993, Quinlan put forward the C4.5 algorithm [16] on the basis of ID3 algorithm where the information Gain rate is applied to select and evaluate the attribute to solve the unfair problem when it comes to the choice of attributes. In addition, the continuous values can be discretized by dichotomy, expanding the application scope of the decision tree. However, C4.5 algorithm is not suitable to process large data because of its constant need to scan and sort data. The C5.0 algorithm [17] is an improvement of the C4.5 algorithm which not only inherits the advantages of the C4.5 algorithm, but also improves the operating efficiency and expanding the memory, so that the operating speed is Greatly increased and C5.0 algorithm is widely used in the classification of large sample. This paper establishes a long-term stock selection model by using C5.0 algorithm to construct the mapping between stock financial index and individual stock return rate.

C5.0 decision tree uses information Gain ratio to select splitting attributes. The attribute is more likely be selected as the splitting attribute if the information Gain rate is larger, fundamentally speaking, it means that the decline extent of information entropy can serve as a basis of feature selection, and the decline of information entropy represents the decline of uncertainty. To sum up, the main idea of the C5.0 decision tree is to construct a rule set with the fastest decline speed of information entropy to Guarantee the information uncertainty of data classified in the same Group is zero.

Let X be the classification of the original data after judging the attributes, so the entropy of X can be expressed as:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

And $p_i = p(X = x_i), i = 1, 2, \cdots, n$ represents the possibility that the sample belongs to the category i.

Knowing the information entropy of category X, then the conditional entropy of category X classified by attribute A can be represented as $H(X|A)$. While the combined probability

distribution of random variable (X, A) is:

$$P(X = x_i, A = A_j) = p_{ij}, i = 1, 2, \cdots, n; j = 1, 2, \cdots k$$

The conditional entropy $H(X \mid A)$ represents the uncertainty of category X under the condition that the attribute A is known, which can be defined as when attribute A is known, the mathematical expectation of entropy of conditional distribution of X to attribute A :

$$H(X \mid A) = \sum_{j=1}^{K} \frac{|A_j|}{|A|} H(X \mid A = A_j) = -\sum_{j=1}^{K} \frac{|A_i|}{|A|} \sum_{i=1}^{N} \frac{|X_{jn}|}{|X_j|} \log 2 \frac{|X_{jn}|}{|X_j|}$$

The information Gain of attribute A to the classification X can be expressed as $g(X, A) = H(X) - H(X \mid A)$, showing the extent of the decrease of information uncertainty of category X when the attribute A is known. On the basis of information Gain, C4.5 defines the splitting information entropy $H_A(X)$ of the classification X with respect to the attribute A as:

$$H_A(X) = -\sum_{j=1}^{K} \frac{|X_j|}{|X|} \log 2 \frac{|X_j|}{|X|}$$

And the information Gain ratio [18] is defined as $gain\_ratio(X, A) = \frac{g(X, A)}{H_A(X)}$

It is obvious that the width of "splitting information" is negatively related with the information Gain ratio. Thus the fairness of the attribute selection was improved Greatly by avoid the shortcoming of choosing the attribute with large value.

Furthermore, the Growth of decision tree is a process of continuously classification of the samples. How to accurately select the best segmentation point in the numerous attribute values becomes a key to improve the accuracy of classification. In C5.0 decision tree, there are different methods to determine the best split value for continuous and discrete attributes. Discrete data can just be divided into different Groups according to the value. Continuous data needs to be sorted in ascending or descending order firstly, and then divided into two sub data sets whose information Gain ratio is calculated. The segmentation point is the threshold value which makes the information Gain ratio largest.

After the comparison and analysis of many different types of algorithms of decision tree, it is easy to observe that C5.0 algorithm is superior compared with other decision tree algorithms no matter in classification accuracy or operation speed. Therefore, this paper uses C5.0 decision tree to build the stock selection model to analyze the relationship between stock financial index and long-term stock return rate.

5. Empirical Test of Stock Selection Model Based on Decision Tree 5.0 Algorithm

The stock data from the second and third quarter of 2017 is used to train the stock selection model. The yield rate is Generally viewed as a lagging factor, so the stock factors in the second quarter corresponds to the yield of the third quarter, and the stock factors of the third quarter corresponds the yield of the fourth quarter. In this paper, samples are divided into training samples and test samples. 70% of data is selected randomly as training sample and the rest is as test sample. The samples of stock can be divided into 6 Grades according to the rate of return which are the excellent, very Good, Good, medium, pass, and fail.

Table 5. *Stocks classification grades.*

Stock Grades	Stock Return
Excellent	>40%
Very Good	(30%,40%]
Good	(20%,30%]
Medium	(10%,20%]
Pass	(0,10%]
Fail	≤0

Table 6. *The classification and classification rate of total samples, training samples and test samples.*

	Number/ Rate of excellent stocks	Number/ Rate of very Good stocks	Number/ Rate of Good stocks	Number/ Rate of Medium stock	Number/ Rate of Pass stocks	Number/ Rate of Fail stocks	Number/ The Classification Rate
Total Samples (1882)	26/1.4%	41/2.2%	76/4.0%	194/10.6%	376/19.9%	1169/62%	1:2:4:11:20:62
Training Samples (1326)	24/1.8%	28/2.1%	57/4.3%	140/10.6%	263/19.8%	814/61.4%	2:2:4:11:20:61
Test Samples (556)	2/0.35%	13/2.3%	19/3.4%	54/9.7%	113/20.3%	355/63.8%	0.5:2:3.5:10:20:64

Compare the distribution of the total sample with that of training sample, finding two distributions are practically the same, which indicates that the training samples can reflect the overall distributions of the total sample. The data mining software SPSS Clementine is used to build C5.0 decision tree model.

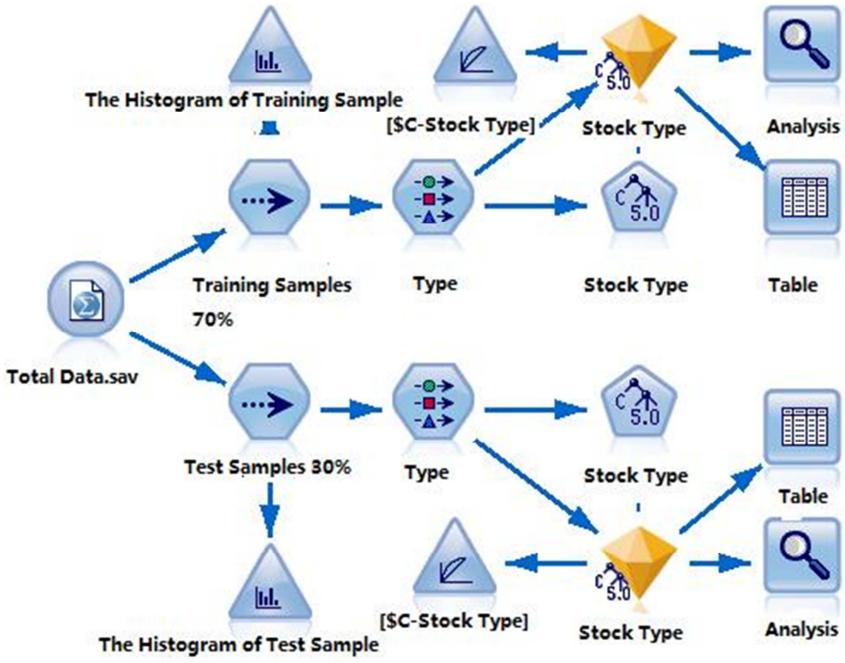

Figure 10. *Flow chart of C5.0 decision tree stock selection model*

The results of classification and prediction of training samples and test samples are shown in the following table:

Table 7. *Classification and prediction of training samples.*

		Predicated Classification						Total	Accuracy Rate	Error Rate
		Excellent	Very Good	Good	Medium	Pass	Fail			
Real Grade	Excellent	0	0	0	1	1	22	24	0	100%
	Very Good	0	0	0	0	1	27	28	0	100%
	Good	0	0	6	1	5	45	57	10.53%	89.47%
	Medium	0	0	0	42	6	92	140	30%	70%
	Pass	0	0	0	0	136	127	263	51.7%	48.3%
	Fail	0	0	0	0	8	806	814	99%	1%
Total								1326	74.66%	25.34%

Table 8. *Classification and Prediction of Test Samples.*

		Predicated Classification						Total	Accuracy Rate	Error Rate
		Excellent	Very Good	Good	Medium	Pass	Fail			
Real Grade	Excellent	0	0	0	0	0	2	2	0	100%
	Very Good	0	0	0	0	0	13	13	0	100%
	Good	0	0	0	0	6	13	19	0	100%
	Medium	0	0	0	6	2	46	54	11.1%	88.9%
	Pass	0	0	0	0	49	64	113	43.4%	56.6%
	Fail	0	0	0	0	6	349	355	98.3%	1.7%
Total								556	72.66%	27.34%

The classification accuracy of the training samples, which is 74.66%, is slightly higher than that of test samples, which is 72.66%. However, both the classification accuracy rates of training samples and test samples showed a decreasing trend from failing Grades to excellent Grades. The classification accuracy of 98% is the highest in the fail category, followed by the pass category, the medium category and the Good category subsequently. The classification accuracy rate of the excellent Group and the very Group is the lowest. The correct rate of classification of pass Group is 51.7% in the training sample while 43.4% in the test sample; the correct rate in the medium category is 30% in the training sample while 11.1% in the test sample; the correct rate of Good category is 10.53% in the training sample. If stocks whose rate of return is less than 0 is defined as "bad" stocks, the stock whose rate of return is higher than 0 is defined as "good" stocks, then it can be clearly concluded from the experiment that the error rate of classifying "good" stock into "bad" stock is much higher than the error rate of classifying "bad" stock as "good" stock. The reason for this phenomenon may be that the number of

"failed" stocks is much Greater than other Groups, occupying large weight ratio and leading to the whole samples tend to be classified as low Grades.

From the perspective of risk aversion, the model effectively reduces the risk cost because the majority of the stocks with poor performance are classified correctly and failure of investment is avoided. But the results of decision tree model may not be satisfied for some radical investors because the model also classifies many outstanding stocks into lower Grades. Finally, the error rate of prediction training sample is very close to that of test sample, indicating that the model is stable, and proving that the same effect will appear on the other new data set.

6. Conclusion

In the process of establishing the stock selection model, this paper firstly selects the industry with excellent development prospects using the industry rotation theory and regards the Gathered stocks in these industries as the initial stock pool. Then the comprehensive multi-factor stock selection is built which combines the factor analysis with decision tree to quantify the relationship between the stock index system and the rate of return. It is observed from the results of the model that cash flow indicators, Growth indicators, technological indicators and operating indicators reflect market information to some extent. Therefore, the classification obtained from the stock selection model is scientific and can be taken into account in investment analysis. All in all, using decision tree to evaluate stock index and classify stocks can effectively Guide investors to inhibit impulsiveness and avoid risks along with invest rationally.

Acknowledgements

This paper is supported by the Natural Science Foundation of Guangdong Province, China under Grant 2017A030313435.

References

[1] Wang Dong, Wu Wen-feng, and Aetna School of Management. "Application of Support Vector Machines Regression in Prediction Shanghai Stock Composite Index." Wuhan University Journal of Natural Sciences8.4(2003):1126-1130.

[2] Panigrahi, S. S., and J. K. Mantri. "Epsilon-SVR and decision tree for stock market forecasting." International Conference on Green Computing and Internet of Things IEEE, 2016:761-766.

[3] Panigrahi, S. S., and J. K. Mantri. "Epsilon-SVR and decision tree for stock market forecasting." International Conference on Green Computing and Internet of Things IEEE, 2016:761-766.

[4] Barak, Sasan, A. Arjmand, and S. Ortobelli. "Fusion of multiple diverse predictors in stock market." Information Fusion 36(2017):90-102.

[5] Wei Xiong, “Application of Decision Tree Algorithm in Stock Analysis and Prediction[J].” Computer Knowledge and Technology (academic exchange), 2.9(2007):764-765.

[6] Huang Lingqin, The Application of Data Mining in Stock Analysis and Prediction[D]. Dalian University of Technology, 2008,12.

[7] Zhang Jingyi, Empirical Research on Stock Investment Based on Data Mining Technology[D]. Chongqing University, 2013,04.

[8] Tao Yuyu, Application of Decision Tree and Neural Network in Stock Classification and Forecasting[D]. Hangzhou Dianzi University, 2013,10.

[9] Haung Ling, Hu Yang, “Stock Data Mining Based on C4.5 Decision Tree[J].” Computer and Modernization (Periodical)10(2015):21-24.

[10] Huang Yue, Analysis of Stock Selection Based on Data Mining Technology[D]. Beijing Foreign Studies University, 2017.06.

[11] Shen Jinrong, Stepwise Regression Algorithm Based on Decision Tree and its Application in Stock Prediction[D], Guangdong University of Technology, 2017,06.

[12] Zhang Huiheng, “On the theory of industrial life cycle[J].” Finance and Trade Research15.6(2004):7-11.

[13] Hu Xiaomei, “The Application of Industry Analysis in Securities Investment Management[J].” Heilongjiang's Foreign Trade and Economic Trade 5(2007):94-96.

[14] Wang Huilin, Stock Selection Strategy and Product Design of Medical Theme Private Equity Fund[D], Nanjing University, 2017.

[15] Quinlan, R. "Introduction of Decision Trees." Machine Learning1.1(1986):81-106.

[16] Quinlan, J. Ross.” C4.5: programs for machine learning.” Morgan Kaufmann Publishers Inc. 1992.

[17] Quinlan J R. “Bagging, boosting, and C4.5[C].” Proc of 14th National Conference on Artificial Intelligence, Portland, Oregon, 1996:725-730.

[18] Quinlan, J. R. "Simplifying decision trees." International Journal of Man-Machine Studies 27.3(1987):221-234.