ELSEVIER

Expert Systems with Applications 31 (2006) 270–274

Expert Systems with Applications

www.elsevier.com/locate/eswa

An effective application of decision tree to stock trading

Muh-Cherng Wu *, Sheng-Yu Lin, Chia-Hsin Lin

Department of Industrial Engineering and Management, National Chiao Tung Uninversity, 1001 Dah-Hsei Road, Hsin-Chu 300, Taiwan, ROC

Abstract

This paper presents a stock trading method by combining the filter rule and the decision tree technique. The filter rule, having been widely used by investors, is used to generate candidate trading points. These points are subsequently clustered and screened by the application of a decision tree algorithm C4.5. Compared to previous literature that applied such a combination technique, this research is distinct in incorporating the future information into the criteria for clustering the trading points. Taiwan and NASDAQ stock markets are used to justify the proposed method. Experiment results show that the proposed trading method outperforms both the filter rule and the previous method.

Keywords: Decision tree; Stock trading; Filter rule

1. Introduction

In a stock market, how to find right stocks and right timing to buy has been of great interest to investors. To achieve this objective, some research used the techniques of *technical analysis*, in which trading rules were developed based on the historical data of stock trading price and volume (Alexander, 1961; 1964; Bessembinder & Chan, 1995; Fama & Blume, 1966; Huang, 1995; Sweeney, 1988; 1990; Szakmary, Davidson, & Schwarz, 1999). Some other research used the techniques of *fundamental analysis*, where trading rules are developed based on the information associated with macro-economics, industry, and company (Al-Debie & Walker, 1999; Lev & Thiagarajan, 1993).

Among the methods of the technical analysis, the technique—*filter rule* has been widely used. The idea of the filter rule is to buy when the stock price rises $k\%$ above its past local low and sell when it falls $k\%$ from its past local high. Alexander (1961) pioneered the technique of the filter rule. He used the data of Dow-Jones Industrials from 1897 to 1927 and the Standard and Poor's Industrials from 1929 to 1959 and found that the application of the filter rule would yield excess returns. However, in a further study also conducted by Alexander (1964), the application of the filter rule may not yield excess returns. The effects of the filter rule, empirically tested at various stock markets and at different time horizons, have been studied (Fama & Blume, 1966; Huang, 1995; Sweeney, 1988; 1990; Szakmary et al., 1999). Research results show that the filter rule may yield excess returns at some stock markets but may not be so at some others.

Lin (2004) proposed a method to modify the filter rule by incorporating three decision variables associated with fundamental analysis. An empirical test, using the stocks of electronics companies in Taiwan, showed her method outperforms the filter rule.

However, in Lin's work, the criteria for clustering trading points involved only the past information; the future information was not considered at all. This research aims to improve the filter rule and Lin's study by considering both the past and the future information in clustering the trading points. The future information is modelled by the price of stock index futures. We use the data of Taiwan stock market and that of NASDAQ to carry out empirical tests. Test results show that the proposed method outperforms both Lin's method and the filter rule in the two stock markets.

The remainder of this paper is organized as follows. Section 2 describes the four variables used to cluster the trading points by decision tree. Section 3 briefly explains the decision tree algorithm C4.5 (Quinlan, 1992) used in this research. Section 4 presents the results of empirical tests. Concluding remarks are in Section 5.

2. Stock trading method

The proposed stock trading method, a modification of Lin's technique (Lin, 2004), first uses the filter rule to find a set of candidate trading points. Then, a decision tree algorithm C4.5 is used to cluster the set of trading points. The criteria for the

* Corresponding author. Tel.: +886 35 731 913; fax: +886 35 720 610.
E-mail address: mcwu@cc.nctu.edu.tw (M.-C. Wu).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.026

clustering involve four variables. This section presents the filter rule, the four variables, the stock trading assumptions, and the metric for justifying stock investment.

2.1. Filter rule and enhancement

The trading method of the filter rule is explained below (Alexander, 1961). Let MA(n) represent the n days moving average of a stock price. If MA(n) moves up at least k% from its past local low, buy and hold the stock until it moves down at least k% from its past local high, at which time simultaneously sell the stock.

The filter rule has two parameters n and k, the values of which will have an effect on the performance of the file rule. The lower is the value of n, the more sensitive is MA(n) because noise signals may impose a significant impact on MA(n). The lower is the value of k, the more is the number of trading which leads to higher trading cost. To effectively use the filter rule, we have to find an optimal set of (n, k) in advance.

The application of filter rule may generate a great number of buying points. Each of these trading points may vary sparsely in its return. This research aims to develop a mechanism to screen these trading points in order to find a set of *effective buying points* that has higher return than the original set in average. We propose the following trading rule. Whenever an *effective buying point* appears, buy the stock; sell the stock whenever a selling signal appears. The proposed method differs with the filter rule in buying signal but is the same with the filter rule in selling signal.

2.2. Variables for identifying effective buying points

This research uses four variables to clarify whether a buying point is an effective one. These four variables are associated with the following factors: (1) money supply, (2) inflation rate, (3) the billings (or revenues) of the upper stream entities in the industry of interest, and (4) the price of stock index futures. The first three variables refer to the past information, while the last one refers to the future information.

Previous literature has revealed that the money supply is positively correlated with the stock price (Friedman, 1988). Let GM(i) represent the money supply growth rate of month i, compared to the same month in the last year. Whenever a buying point appears at month i, we take a linear regression for the six data GM(i), GM($i-1$),…,GM($i-5$) and compute a slope. The slope is taken as the first variable for clarifying effective buying points.

This research uses CPI (consumer price index) to represent the inflation rate. Previous studies reveal that CPI has an effect on stock price (Fama, 1981; Hu & Willett, 2000). A slightly increase in CPI will help the growth of economy and subsequently increase the stock price. However, a high increase in CPI will discourage the growth of economy, which subsequently leads to the decrease of the stock price. Let GC(i) represent the CPI growth rate of month i, compared to the same month in the last year. Whenever a buying point appears at month i, we take a linear regression for the six data GC(i), GC($i-1$),…,GC($i-5$) and compute a slope. The slope is taken as the second variable for clarifying effective buying points.

In a supply chain, the upper stream entities have impacts on the down stream entities. For example, the semiconductor industry is the upper stream of an electronics supply chain. Let TU(i) represent the total billings of the upper stream industry in month i. Whenever a buying point appears at month i, we take a linear regression for the six data TU(i), TU($i-1$),…,TU($i-5$) and compute a slope. The slope is taken as the third variable for clarifying effective buying points.

The fourth variable refers to the price of stock index futures. In a typical stock market, five months of futures can be traded at every trading day, which involve the spot month (or current month), the next calendar month, and the next three quarter months. Let GF(j), GF($j-1$),…,GF($j-4$) represent the five feasible trade of futures at day j, at which time a buying point appears. We take a linear regression of the five data. The computed slope of the regression is taken as the fourth variable for clarifying effective buying points.

In summary, each of the variables for clarifying effective buying points is modelled by a slope of a line. To clarify an effective buying point is to define a *valid region* for each variable; that is, defining the upper bound and lower bound for the value of each variable. Whenever a buying point appears, if all the four slopes fall within the valid regions, then this point is regarded as an effective point; otherwise it is not effective and should be discarded.

2.3. Assumptions and performance metric

The scenario assumptions are essentially adopted from the present regulations of Taiwan stock market. The commission for each trading transaction, whether buying or selling, is 0.1425%. The trading tax, applicable for selling only, is 0.3%. The trading policy and assumptions are defined as follows. Whenever an effective buying signal appears, we can always buy the stock at the price of next day's closing price. Suppose a stock has been purchased, we will not buy the stock further until the presently hold stock has been sold. At the end of the observation horizon, all stocks in hand have to be sold.

The performance metric is the averaged compound annual rate of return (ACARR), which is computed as follows. Let r_i represent the rate of return of i-th buying point, s_i represent the selling price of i-th buying point, b_i represent the buying price of i-th buying point, h represent the rate of commission, and o represent the tax for trading. Then

$$r_i = \frac{(s_i \times (1-h-o) - b_i \times (1+h))}{b_i \times (1+h)}$$

Let $N(i)$ represent the number of trading transactions at year i, r_{it} is the rate of return of t-th transaction at year i. We can compute R_i, the annual rate of return for year i, as follows.

$$R_i = \prod_{t=1}^{N(i)} (1 + r_{it}) - 1$$

The performance metric ACARR, the averaged compound annual rate of return, can be computed below.

$$\text{ACARR} = \sqrt[n]{\prod_{i=1}^{n}(1+R_i)} - 1$$

3. Decision tree algorithm

This research uses a decision tree algorithm C4.5 (Quinlan, 1992) for clustering trading points. The algorithm C4.5, widely used in literature (Sebban, Mokrousov, Rastogi, & Sola, 2002; Viaene, Derrig, Baesens, & Dedene, 2002; Yang, 2004), is a *supervised-clustering* technique for grouping objects. The candidate trading points forms a population of interest. The decision tree algorithm can yield a set of *classification rules* for clustering the population with tolerable error. A vector comprising the four variables stated in Section 2 is used to model each trading point. A cluster is clarified by specifying the feasible range for the value of each variable. Detailed procedure of C4.5 can be referred to (Quinlan, 1992).

Before carrying out the classification by C4.5, we have to explicitly define the criterion for assigning a candidate trading point to one of two classes—a *favourable* and an *unfavourable* class. The criterion is a threshold of investment return, denoted by H. Any candidate trading point i, if its return r_i is less than H, should be assigned to the unfavourable class, otherwise assigned to the favourable class. In the application of algorithm C4.5, different settings of H value will yield different ACARR.

4. Empirical tests

The *electronics stocks* in Taiwan stock market and the *technology stocks* in NASDAQ market are used to justify the proposed trading method. The observation horizon for Taiwan stock market ranges from July 1998 to December 2004, and that for NASDAQ ranges from January 1997 to December 2004. Such a sampling involves 41 Taiwan stocks and 248 NASDAQ stocks. The prices of stocks have been adjusted by removing the effects of stock dividends.

Of the four variables considered in C4.5, the three—money supply, CPI and the price of stock index futures refer to the metrics in the stock market of interest. For the remaining variable, we use semiconductor billings in Americas for NASDAQ, and use semiconductor billings in Asia Pacific for Taiwan stock market.

A data mining software WEKA, freely available on the web, is used in the application of C4.5. The value of H for clarifying the favourable and unfavourable classes is set at nine levels: −9, −6, −3, 0, 1, 2, 3, 6, 9%.

In the application of the filter rule, we first determine an optimal portfolio of parameters (n, k). We test 70 scenarios of (n, k), with $n = 2, 4, 6, 8, 10, 30, 72$ and $k = 1$–10%. Experiment results reveal that the filter rule performs the best at $(n, k) = (10, 10\%)$ both in Taiwan and NASDAQ markets. In the following performance comparison for trading methods, any application of the filter rule refers to just the scenario $(n, k) = (10, 10\%)$.

Table 1 shows the performance of the proposed trading method at various H values for the case of NASDAQ. The proposed method at $H = 3\%$ yields an optimum ACARR = 12.53%, better than Lin's method (ACARR = 8.70%) and the filter rule (ACARR = 5.87%). The table also indicates that an appropriate selection of H value is very important in applying the proposed method. Table 2 shows the returns with $H = 3\%$ at various layers of the decision tree, which depicts that clustering the trading points up to *layer three* will yield the maximum return. The trading rules at layer three reveals that the future information (price of futures) is the most critical variable, the CPI is the second, and semiconductor billing is the third, as shown in Table 3. The table also depicts feasible ranges of the three critical variables, where PF (price of futures) denotes the fourth variable, CPI (inflation rate) denotes the second variable and SEMI (semiconductor billings) denote the third variable.

Table 4 shows the performance of the proposed trading method at various H values for the case of Taiwan. The proposed method at $H = 1$ or 2% yields an optimum ACARR = 13.26%, better than Lin's method (ACARR = 11.50%.) and the filter rule (ACARR = 1.24%). The table indicates that an appropriate selection of H value is very important to obtain a high return in the application of the proposed method. Table 5

Table 1
Performance of the proposed method for the case of NASDAQ

Value of H (%)	ACARR of filter rule (%)	ACARR at different layer of the decision tree				
		1st (%)	2nd (%)	3rd (%)	4th (%)	5th (%)
9	5.87					
6	5.87					
3	5.87	9.73	12.53	7.34		
2	5.87	6.74	7.56	4.08		
1	5.87	8.70	4.53	7.68	11.19	7.37
0	5.87	6.74	7.56	1.98	2.87	
−3	5.87	6.76	7.58	1.97	2.87	
−6	5.87	6.73	7.55	1.98	2.87	
−9	5.87	9.73	10.10	3.33		

Table 2
Results of the decision tree for the case of NASDAQ

	Filter rule	Layer of the decision tree result		
		1st	2nd	3rd
ACAR	5.87%	9.73%	12.53%	7.34%
Number of trading points	5340	4686	2093	750
% of trading points with positive return	38%	40%	47%	57%

Table 3
Feasible ranges of variables at each layer for NASDAQ case

Layer	Feasible ranges of variables at each layer
1st	PFϕ50
2nd	PFϕ50; CPI > −0.000006
3rd	PFϕ50; CPI > −0.000006; SEMI > 0.086

Table 4
Performance of the proposed method for the case of Taiwan

Value of H (%)	ACARR of filter rule (%)	ACARR at different layer of the decision tree					
		1st (%)	2nd (%)	3rd (%)	4th (%)	5th (%)	6th (%)
9	1.24	5.32	9.03	13.04	10.91	6.44	
6	1.24	5.32	9.03	13.04	10.91	4.48	
3	1.24	5.51	8.99	13.00	10.91	4.48	
2	1.24	5.51	9.66	13.26	9.44	4.93	
1	1.24	5.51	9.66	13.26	9.44	4.93	
0	1.24	5.51	9.66	12.98	9.44	4.93	
−3	1.24	5.44	9.10	13.02	9.49	4.93	
−6	1.24	4.95	9.16	12.48	9.44	1.99	4.61
−9	1.24	4.95	9.68	5.70	8.31		

Table 5
Results of the decision tree for the case of Taiwan

	Filter rule	Layer of the decision tree				
		1st	2nd	3rd	4th	5th
ACAR	1.24%	5.51%	9.66%	13.26%	9.44%	4.93%
Number of trading points	623	528	456	344	133	105
% of trading points with positive return	34%	39%	43%	51%	66%	56%

shows the returns with $H=1\%$ at various layers of the decision tree, which depicts that clustering the trading points up to layer three will yield the maximum return. The trading rules at layer three reveals that the future information (price of stock index futures) is the most critical variable and the semiconductor billings is the second, as shown in Table 6. The table also depicts the feasible ranges of the two critical variables, where PF (price of futures) denotes the fourth variable and SEMI (semiconductor billings) denote the third variable.

Empirical tests of the two stock markets indicate that future information is more important than past information. This finding confirms our research hypothesis. Incorporating the price of stock index futures into the criteria indeed improve the performance of the decision tree in clustering trading points.

5. Concluding remarks

This paper presents a stock trading method, which is essentially an enhancement of the filter rule. The buying points generated by the filter rule are further clustered and screened by the application of decision tree. The criteria for the clustering involve four variables, three of which are associated with the past information. The remaining variable is associated with the future information.

Such a trading technique by combining filter rule and decision tree has been used by Lin (2004). However, this research is distinct in two fold. First, the future information for clustering trading points is not considered in Lin's method. Second, we use various H values rather than a single one in the application of the decision tree. Empirical tests show that the two distinctions indeed improve the performance of the decision tree.

Table 6
Feasible ranges of variables at each layer for Taiwan case

Layer	Feasible ranges of variables at each layer
1st	PFϕ43.1
2nd	PFϕ43.1; −0.072(SEMI
3rd	PFϕ43.1; −0.072(SEMIϕ0.19
4th	9.1(PFϕ43.1; −0.072(SEMIϕ0.19
5th	9.1(PFϕ43.1; −0.072(SEMIϕ0.19; CPIϕ−0.55

Two stock markets, Taiwan and NASDAQ, are used to justify the proposed method. Empirical tests reveals that the filter rule performs the best at $(n, k)=(10, 10\%)$ in both the two markets. The proposed trading method outperforms Lin's method, substantially in NASDAQ market and slightly in Taiwan. Our study also confirms that Lin's method outperform the conventional filter rule substantially.

Future extensions of this research involve incorporating some other variables into the criteria for clustering the trading points. These include new variables that reflect future information and those that reflect the impacts of other stock markets to the market of concern.

References

Al-Debie, M., & Walker, M. (1999). Fundamental information analysis: An extension and UK evidence. *Journal of Accounting Research*, *31*(3), 261–280.

Alexander, S. S. (1961). Price movements in speculative markets: Trends or random walks. *Industrial Management Review*, *2*(2), 7–26.

Alexander, S. S. (1964). Price movements in speculative markets—trends or random walks, number 2. *Industrial Management Review*, *5*(000002), 25–46.

Bessembinder, H., & Chan, K. (1995). The profitability of technical trading rules in the Asian stock markets. *Pacific-Basin Finance Journal*, *3*(2–3), 257–284.

Fama, E. F. (1981). Stock returns, real activity, inflation, and money. *The American Economic Review*, *71*(4), 545–565.

Fama, E. F., & Blume, M. E. (1966). Filter rules and stock-market trading. *Journal of Business*, *39*(1), 226–241.

Friedman, M. (1988). Money and the stock market. *Journal of Political Economy*, *96*(2), 221–244.

Hu, X., & Willett, T. D. (2000). The variability of inflation and real stock returns. *Applied Financial Economics*, *10*(6), 655–665.

Huang, Y. S. (1995). The trading performance of filter rules on the Taiwan stock exchange. *Applied Financial Economics*, *5*(6), 391–395.

Lev, B., & Thiagarajan, R. (1993). Fundamental information analysis. *Journal of Accounting Research*, *31*(2), 190–215.

Lin, C. H. (2004). Profitability of a filter trading rule on the Taiwan stock exchange market. Master thesis, Department of Industrial Engineering and Management, National Chiao Tung University.

Quinlan, J. R. (1992). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Sebban, M., Mokrousov, I., Rastogi, N., & Sola, C. (2002). A data-mining approach to spacer oligonucleotide typing of *Mycobacterium tuberculosis*. *ProQuest Biology Journals*, *18*(2), 235–243.

Sweeney, R. J. (1988). Some new filter rule tests: Methods and results. *Journal of Financial and Quantitative Analysis*, *23*(3), 285–300.

Sweeney, R. J. (1990). Evidence on short-term trading strategies. *Journal of Portfolio Management*, *17*(1), 20–26.

Szakmary, A., Davidson, W. N., & Schwarz, T. V. (1999). Filter tests in Nasdaq stocks. *The Financial Review*, *34*(1), 45–70.

Viaene, S., Derrig, R. A., Baesens, B., & Dedene, G. (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance*, *69*(3), 373–421.

Yang, Z. R. (2004). Mining gene expression data based on template theory. *ProQuest Biology Journals*, *20*(16), 2759–2766.