

RELATIONSHIP BETWEEN FUNDAMENTAL VALUES IN THE FINANCIAL MARKET

Rui Liu and Mehdi Zargham

Department of Computer Science
Southern Illinois University Carbondale
Email:mehdi@cs.siu.edu

ABSTRACT

There is a large amount of data and fundamental variables in the stock market. In this work, only seven of these variables are selected and considered. Fundamental variable pairs are constructed from these variables and their visualization figures are built. Then a statistical method is used to find the sub-areas with high frequency. With the help of these high frequency sub-areas, we observe all the visualization figures for every variable pair. From the observations, we can find sub-areas in which the return of stocks is better than the average return of all stocks. Based on these sub-areas, a set of rules is derived using the training data sets from 1993 to 1998 and are tested on the data set from 1999 to 2003. The average of returns of most of these rules is better than the average of returns of all stocks in S&P 500.

Keywords: *Correlation, visualization, fundamental variable, financial market.*

1. INTRODUCTION

Stocks are among the most talked about and most popular investment opportunities available. The trend of the stock market is hard to predict, but it does not mean the prices of stocks are randomly generated. Certainly, there are hidden rules associated with the stock performance in the financial market. The most famous rules for selecting stocks are attributed to Benjamin Graham [1]. These rules performed well before the year 1976. However, the stock market has had dramatic changes since 1976, making these rules unsatisfactory in today's stock markets. Some software, such as PORSEL (PORtfolio SElection system), has already been developed to select and evaluate stocks based on the derived rules [2].

In general, many techniques can be applied in the stock market. These techniques are used in stock prediction and stock performance evaluation. They may obtain some results under certain condition or get good performances for some particular years. However, these methods are lacks of generality. The decision tree method provides us with a way to generate "If-Then" rules from data, and has been applied in the stock market to extract rules for stock prediction based on decision tree C4.5 [3]-[4]. The authors of [5] also applied the decision tree method to extract rules for US IPO data. The limitation of this method is that it cannot deal with continuous attributes efficiently. To solve this problem, fuzzy decision tree method was applied by Zheng [6] to derive rules for stock selection. The fuzzy decision tree is an attractive method used to predict stock selection; one disadvantage of it may be the lack of dynamic fitting because the fuzzy sets used in fuzzy decision tree cannot be changed after creation. Machine learning approaches are also applied for stock prediction and rule extraction. In [7], a reinforcement learning algorithm was introduced for stock price prediction. The authors of [8] proposed a fuzzy support vector machines regression method to forecast the stock composite index. Another very important class of approaches used frequently in securities analysis is the artificial neural networks (ANNs).

Due to its flexibility and robustness, it has become very important in making stock market predictions and is mostly implemented in forecasting stock prices and returns. In [9], neural networks were used for modeling financial time series. In [10]-[11] the authors proposed neural network approaches to predict the stock market. The authors of [12] used a neural network for rule extraction for analyzing dividend events. In addition, some researchers tend to include novel factors in the learning process for neural networks. The authors of [13] incorporated prior knowledge to improve the performance of stock market prediction. In [14], the authors integrated the rule-based technique and ANN to predict the direction of the S&P 500 stock index futures on a daily basis. In [15], the authors proposed an ANN stock selection system to select stocks that are top performers from the market and to avoid selecting under performers. A fuzzy neural network was also proposed in [16] to extract fuzzy rules. In general, ANNs have a lot of advantages and can be easily combined with other AI approaches in stock prediction. However, they are hard to train and may not produce a good solution.

The methods discussed above may have satisfactory results within some period when applied to the stock market. However, each of them has its own limitation. In this study, first we visualize the securities data to get nonlinear correlations between fundamental variable pairs; from the visualization, we can extract rules which can be used in the selection of stock. Lastly, the performances of the rules are tested. The stocks are selected based on the rules and see whether they have a higher return than the average return of all stocks in S&P 500.

The rest of the paper is organized as follows: In section 2 we propose a methodology to visualize data sets into data points and provide a criterion to color these data points. In section 3 we study the visualization figures. Then we derive rules based on visualization results and test performance of these rules. Conclusions and recommended study are presented in section 4.

2. DATA VISUALIAZTION

In this research, all data samples are constructed from the S&P 500 database (The database is accessed by using the software S&P Research and Insight). The fundamental variables of the securities which could be directly collected from the S&P 500 database are MonthPrice, Book Value Per Share (BKVLPS), Current Ratio (CR), Dividend Yield (DXYDF), Price to Book Ratio (MKBK), Price to Earning Ratio Monthly (PEM), % Earning Growth (EG) and Earning Stability (ES).

2.1 Data preparation

The values of these fundamental variables are prepared to construct data sets which will be used later.

2.1.1 Return and Return Class

MonthPrice is used to generate two attributes about return values: *Return* and *Return Class*.

Return is used as the main attribute for data samples (a data sample corresponds to a particular company). The definition of *Return*, in a particular year, is the MonthPrice of December in that year minus the MonthPrice of December in the previous year and then divided by the MonthPrice of December in the previous year. This attribute is a parameter for us to see whether stocks increase or decrease in value and to what extent they can be. Note all the data samples in this research must have *Return* values.

In order to classify *Return* values in different groups, we transform them into another attribute called *Return Class*. Unlike *Return*, whose value could be any real number, *Return Class* just has five integer values which are 0, 1, 2, 3, and 4. Each of the values corresponds to a number of *Return* values in different ranges and has different meanings. When the Return value is less than 0, its Return Class is 0; when the Return value is bigger than or equal to 0 bur less than 0.1, its Return Class is 1; similarly, when the Return value is between 0.1 to 0.3, its Return Class is 2; when the Return value is between 0.3 to 1, its Return Class is 3; when the Return value is bigger than or equal to 1, its Return Class is 4. For example, when the *Return* value of a sample is 0.5,

which is in the range of 0.3 and 1, the corresponding *Return Class* of this sample is 3.

2.1.2 Data sets construction

Except for *MonthPrice*, all the other fundamental variables are required to be annual variables in the data set. An annual variable means that this variable of one data sample has only one value per year. Some fundamental variables such as *BKVLPS*, *CR*, *DVYDF*, *EG* and *ES* are annual variables, so their values will be directly used for each year. Other variables like *MKBK* and *PEM* are monthly variables. We will choose the value in December as the value of the variable in that year to change them into annual variables. In this study, we construct data sets that include fundamental variables of a certain year and the *Return* (as well as *Return Class*) values of next year. For each year, we have a corresponding data set to represent securities data.

2.2 Data sample visualization and variable range determination

Since some information of the data samples cannot be shown if we study a single fundamental variable only, a better way to find the relationship of data samples is using those variables as *variable pairs*. From previous discussion we know that variable *MonthPrice* is used to calculate *Return* and *Return Class* and cannot be used for variable pairs. Therefore, the other seven fundamental variables could be randomly chosen as pairs. The number of pairs, clearly, is twenty one (C_7^2). We associate each pair with a unique *pair key* (a number to identify which variables are included in this pair).

Each pair has exactly two fundamental variables; we will let one of them be the x axis and the other the y axis. Therefore, every data sample has a value on the x axis and a value on the y axis which form a point on the two-dimensional visualization figure. Note, if a data sample does not have either of these two values in a certain pair, this sample is not considered in the visualization figure or in later research. Note that it does not matter which variable in the variable pairs we should choose to appear on the x axis and which one should be on the y axis because it will not affect the distribution of the data points.

After visualization of all the data samples from our data sets, the points are distributed in two-dimensional visualization figures. Some points distributed far away may cause a large area in which the density of the points is very small (large area has only small number of points). In our study, we pay more attention to the area in which most of the points appear. So this area with high point density is our objective to study and we would like to set the range of it. The ranges of variables are not fixed. They are subjective to different researchers. The number of the points that are out of this range is small compared to those in this range, so we will not consider them in our later approaches.

2.3 Data point density calculation

To distinguish density of the points accurately, we divide each area into sub-areas and classify them into a different density level based on number of points in them. Then according to the different density levels of the sub-areas, we give different colors to the points in these sub-areas.

2.3.1 Sub-area generation

According to the range of each fundamental variable pair, we can find a corresponding area. We divide each area average into 100 sub-areas and associate each sub-area with a unique integer number from 1 to 100 and we call each number *sub-area ID*. In this study, the *ID* of leftmost downward sub-area is 1 and the *ID* number increases in the same row from left to right (also from downward to upward).

2.3.2 Density levels calculation

The density levels for different fundamental variable pairs are obviously different. The following are the steps used to get the density levels for the particular fundamental variable pair based on a certain data set.

Step 1. From the data set, we choose the *Return Class* and two fundamental variables in the particular pair to form a useful data set S . The other columns of the data set which will not be used are ignored at this time.

Step 2. We divide this useful data set S into five subsets S_0, S_1, S_2, S_3 and S_4 according to the values of *Return Class*. In subsets S_0 , the values of *Return Class* of all the data samples are 0. Similarly, the *Return Class* values in subsets S_1, S_2, S_3 and S_4 are 1, 2, 3 and 4 respectively.

Step 3. In subset S_0 , we visualize each data sample whose values are within the variable range and count the number of data points in every sub-area. Assume M_0 is the maximum number of data points per sub-area.

Step 4. For subsets S_1, S_2, S_3 and S_4 , we repeat Step 3 and get the maximum number of data points per sub-area, namely M_1, M_2, M_3 and M_4 . To unify the density level of all kinds of *Return Class*, assume M is the maximum number of data points per sub-area for the whole data set S , that is $M = \max\{M_0, M_1, M_2, M_3, M_4\}$. In this study, there are five density levels where each of them is with length $len = M/5$.

Step 5. Suppose the number of points in a sub-area n ($n \in [1, \dots, 100]$) is $P(n)$, the density level of one sub-area n ($n \in [1, \dots, 100]$) is $L(n)$, we have: if $0 \leq P(n) < len$, $L(n) = 1$; if $len \leq P(n) < 2len$, $L(n) = 2$; if $2len \leq P(n) < 3len$, $L(n) = 3$; if $3len \leq P(n) < 4len$, $L(n) = 4$ and if $P(n) \geq 4len$, $L(n) = 5$.

2.4 An example of visualization

In this subsection, there is an example to show how to visualize the data samples and color the data points. Let's take the data set from the year 1994 as an example. At this time, we would like to pay attention to the variable pair CR and MKBK. The range of CR is $[0, 10]$, and the range of MKBK is $[0, 20]$. Here, CR is chosen as the x-axis and MKBK as the y-axis. First, we choose the useful information from the whole data set to build the useful data set S . According to the *Return Class* values, every data sample could be put in its corresponding subset. For instance, if the *Return Class* of a sample is 2, this sample will be put in S_2 . Next, we use MATLAB to plot the data samples from different subsets in different figures as the left figure in Figure 1:

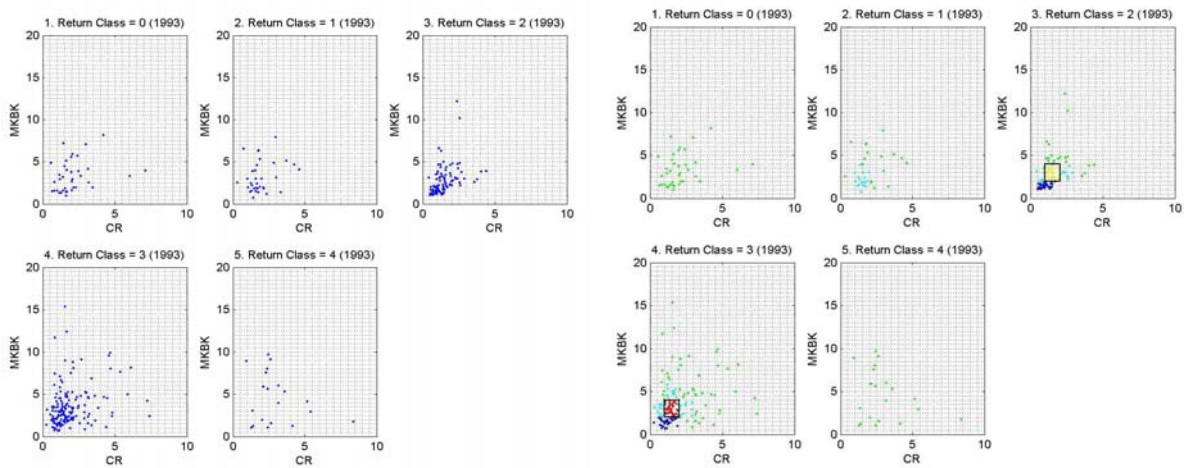


Figure 1. No-colorful visualization and colorful visualization for the variable pair CR and MKBK (1994)

Each subfigure in the left figure in Figure 1 shows the data samples with particular value of *Return Class*. Also, it is easy to count the number of points of each sub-area using MATLAB. Another advantage of using MATLAB is that it avoids the problem of miscounting if two points are overlapping. The maximum number of points per sub-area from all the 500 sub-areas (each subfigure contains 100 sub-areas), M , is 36. This number is calculated using MATLAB. According to the maximum number of points per sub-area M , which is equal to 36, the length of each density interval is $36/5=7.2$. The Table 1 is an example of the density levels' definition, showing relationship of different density levels, the range of the number of data points in the sub-area and the color of the data points in that area.

Table 1. An example of density levels definition ($M = 36$)

Density Level	The Range of Point Numbers	The Color of the Points
5	(28.8, 36]	Red
4	(21.6, 28.8]	Yellow
3	(14.4, 21.6]	Blue
2	(7.2, 14.4]	Cyan
1	[0, 7.2]	Green

Based on Table 1, we can redraw the left figure in Figure 1 into its colorful version, as shown in the right figure of Figure 1 in which we can get more information about the distribution of data points. Take the sub-area 12 (as shown in black rectangular in the right figure of Figure 1) an example; it is very clear that this area has more data points when the Return Class is 3 than when the Return Class is 2 because the red points denote higher density level than yellow points. However, it is difficult for us to see that in the left figure of Figure 1.

This example shows the visualization of one variable pair (CR and MKBK) for the data set 1994. Similarly, we could visualize the other twenty variable pairs. That is, we could generate 21 figures for every data set if there are enough variable values. Totally, from the 6 data sets (1993-1998) we generated more than one hundred figures like the right figure in Figure 1.

3. RULE EXTRACTION AND VERIFICATION

3.1 Calculating the sub-area frequencies of each variable pair

There are 100 sub-areas in visualization figure of a certain variable pair. These sub-areas have their values of density levels for each *Return Class* in a certain year. And then, we want to know the information which the sub-area corresponds to high density levels (3, 4 and 5) for each *Return Class*. So for each *Return Class* and each high density level, we let *Sub-area Set* to include all the sub-areas which correspond to them. There are a total of fifteen *Sub-area Sets* for each variable pair (five values of Return Class multiplied by three density levels). For different year (different data set), the fifteen sub-area sets are different for the same variable pair.

For each variable pair, we want to find how many times (the *frequency*) each sub-area corresponds to a specific *Return Class* and specific density level of all the data sets. To do so, we unify the sub-area sets with the same *Return Class* and the same density level. We call this new union set a *statistic set*. Note that we just take the union of sub-area sets, so there are 15 *statistic sets* for each variable pair. For each *statistic set*, we can calculate the *frequency* $f(ID)$ of each sub-area ID in each set (how many times that sub-area ID appears in that set). For example, if *statistic set* is {2, 2, 12} for a particular *Return Class* and density level, $f(2)=2$ and $f(12)=1$ (sub-area 2 appears twice and sub-area 12 appear once in this statistic set). Data sets from 1993 to 1998 are used to calculate the sub-area frequencies of each variable pair. The sub-areas with high frequencies are found from the statistic results. Compared to sub-areas with low frequencies, these sub-areas are more likely to provide more general information.

3.2 Observation results from visualization figures

Considering these high frequency sub-areas, we review the visualization result for each variable pair and attempt to find hidden rules.

Initially, we want to find some variable pairs who have different sub-areas for good and bad return stocks. For some particular year, the sub-areas for good return are different from the sub-areas with bad return. However, this phenomenon does not occur in all of the other years for the same variable pair. The results from this kind of observation can be seen as an “accident” but cannot be developed as a general rule for stock prediction.

Although we cannot find a sub-area in which the data samples always have a good or bad return in any variable pair, we notice that time is a very important factor that could affect the return of stocks. For some variable pairs, most of the data points may be plotted in the first subfigure (Return Class = 0) in one year, while many data points may be plotted in the fourth subfigure (Return Class = 3) in another year. Most stocks had a high or low return because of the great change in the economy. This situation inspires us to compare the individual stock's performance of each sub-area with the performance of all the stocks. We calculate the average return of all stocks and the average return of stocks in each sub-area. Then we put the result in the form of a table which is convenient for us to compare for each variable pair.

After studying all of these tables, we notice that some sub-areas work well in all 6 years. Based on these special sub-areas, we can derive some rules in "IF-THEN" form.

In this study, we generate a total of four rules which are as follows:

1. If BKVLPS is within [0, 5] and CR is within [1, 2], then the return is better than the average.
2. If BKVLPS is within [0, 5] and DVYDF is within [0, 1], then the return is better than the average.
3. If BKVLPS is within [0, 5] and MKBK is within [4, 6], then the return is better than the average.
4. If DVYDF is within [0, 1] and MKBK is within [4, 6], then the return is better than the average.

Here, "the return is better than the average" means the average return of the stocks within both of the ranges in IF condition is higher than or equal to the average return of all the stocks.

3.3 Rules performance

In the previous section, we derived rules from data sets between 1993 and 1998. In this section, we show the performance of these rules using data sets from 1999 to 2003. Note that here fundamental variables are from data sets between 1999 and 2003 and the return values are from 2000 to 2004.

According to certain rules, we chose stocks within the ranges of the corresponding fundamental variables; then we can see the average of *Return* values for these stocks in the next year. The performances of all rules are shown in Table 2:

Table 2. The performance of all rules

	1999	2000	2001	2002	2003
Average	0.192	0.008	-0.169	0.536	0.162
Rule 1	0.246	0.048	-0.136	0.610	0.166
Rule 2	0.287	0.039	-0.145	0.673	0.178
Rule 3	0.213	0.153	-0.052	0.671	0.179
Rule 4	0.724	0.217	-0.218	0.638	0.281

In Table 2, the second row shows the average of the returns of all the stocks in S&P 500. The third through sixth rows show the average of the returns of the stocks selected by each rule. From the Table 2, the average of the returns of most of the rules is better than the average of the returns of all stocks in the S&P 500 except the rule 4 in year 2001. We can say that most of the rules perform well. Furthermore, the performance of some of the rules (rule 2 in the year 1999 and the year 2002) is sometimes very good because the average returns from those rules are almost 50% larger than the average of returns of all the stocks in S&P 500.

4. CONCLUSIONS

The main goal of this research was to derive a set of rules which can be used to select stocks with better than average returns. From these rules, we can see the correlations (not linear) between fundamental variables and how they are related with the return of the stocks. In other words, we can find the ranges of these variables within which the corresponding stocks have better than average returns.

In this paper, any two fundamental variables were chosen to become pairs and the results were based on

these variable pairs. Similarly, we can derive rules where multiple variables are chosen in one group. For these kinds of groups, the corresponding multi-dimensional visualization of data becomes more difficult to realize but the methods of statistic sub-areas frequencies are still useful.

5. REFERENCES

- [1] P. Blustein, "Ben Graham's last will and testament", *Forbes*, August 1, pp. 43-45, 1977.
- [2] M. R. Zargham and M.R. Sayeh, "A web-based information system for stock election and evaluation," *International Workshop on Advance Issues of E-Commerce and Web-Based Information Systems*, pp. 81, 1999.
- [3] N. Ren, "Rule Extraction for Security Analysis Based on Decision Tree Classification Model", Master's Thesis, Southern Illinois University Carbondale, May 2004.
- [4] J. R. Quinlan, *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann, 1993.
- [5] R. Mitsdorffer, J. Diederich and C. Tan, "Rule extraction from technology IPOs in the US stock market," *Proc. 9th Int. Conf. Neural Information Processing*, vol. 5, pp. 2328-2334, 2002.
- [6] W. Zheng, "Fuzzy decision tree based rule extraction in security analysis," Master's Thesis, Southern Illinois University Carbondale, July 2005.
- [7] J. W. Lee, "Stock price prediction using reinforcement learning," *Proc. IEEE Int. Symp. Industrial Electronics*, vol. 1, pp. 690-695, 2001.
- [8] Y. K. Bao, Z. T. Liu, L. Guo and W. Wang, "Forecasting Stock Composite Index by Fuzzy Support Vector Machines Regression," *Proc. Int. Conf. Machine Learning and Cybernetics*, Vol 6, pp. 3535-3540, Aug. 2005
- [9] R. Sharda and R. B. Patil, "A connectionist approach to time series prediction: an empirical test," in R. R. Trippi, E. Turban, (Eds.), *Neural Networks in Finance and Investing*, pp. 451-464, 1994.
- [10] H. Ahmadi, "Testability of the arbitrage pricing theory by neural networks," *Proc. Int. Conf. Neural Networks*, pp. 385-393, 1990.
- [11] J. H. Choi, M. K. Lee and M. W. Rhee, "Trading S& P 500 stock index futures using a neural network," *Proc. Annual Int. Conf. Artificial Intelligence Applications on Wall Street*, pp. 63-72, 1995.
- [12] M. Dong and X. S. Zhou, "Analyzing dividend events with neural network rule extraction," *Proc. Int. Joint Conf. Neural Networks*, vol. 4, pp. 2854-2859, Jul. 2003.
- [13] K. Kohara, T. Ishikawa, Y. Fukuhara and Y. Nakamura, "Stock price prediction using prior knowledge and neural networks," *Int. J. Intell. Syst. Accounting Finance Manage*, vol. 6, no. 1, pp. 11-22, 1997.
- [14] R. Tsaih, Y. Hsu and C. C. Lai, "Forecasting S& P 500 stock index futures with a hybrid AI system," *Decision Support Syst.*, vol. 23, no. 2, pp. 161-174, 1998.
- [15] T.-S. Quah and B. Srinivasan, "Improving returns on stock investment through neural network selection," *Expert Syst. Appl.*, vol. 17 pp. 295-301, 1999.
- [16] T. Nishina, M. Hagiwara and M. Nakagawa, "Fuzzy inference neural networks which automatically partition a pattern space and extract fuzzy if-then rules," *Proc. IEEE Conf. Computational Intelligence*, vol. 2, pp. 1314-1319, Jun. 1994.