Stock Market Prediction Based on Interrelated Time Series Data

Ryota Kato and Tomoharu Nagao
Graduate School of Environment and Information Sciences,
Yokohama National University
Kanagawa, Japan
Email: ryota@nlab.sogo1.ynu.ac.jp

***Abstract*—In this paper, we propose a stock market prediction method based on interrelated time series data. Though there are a lot of stock market prediction models, there are few models which predict a stock by considering other time series data. Moreover it is difficult to discover which data is interrelated with a predicted stock. Therefore we focus on extracting interrelationships between the predicted stock and various time series data, such as other stocks, world stock market indices, foreign exchanges and oil prices. We test our method for predicting the daily up and down changes in the closing value by using discovered interrelationships, and experimental results show that our methods can predict stock directions well, especially in the manufacturing industry.**

***Keywords*—data mining; stock market prediction; Evolution Strategy.**

I. Introduction

These days, many kinds of data are stored and it is easy to gain access to them. On the other hand we are faced with unmanageability of this data. We still don't know what is the best way to use stored data. Therefore, one of the most important problems is to discover knowledge from these data and make effective use of them. For this reason, studying about data mining or knowledge discovery is required. In recent years, it has been possible to analyze huge amounts of data due to developments in computer technology. There are many studies in this field [1], [2].

Regarding financial areas, there have been many studies of the stock market prediction or stock investment using data mining techniques. Indeed, there are many investors using these techniques all over the world. The important things in this area are to discover effective data and to make effective use of the discovered data.

There are two approaches to stock market predictions. One approach focuses on data on the predicted stock, the other approach focuses on data aside from the predicted stock. Concerning the first approach, there are two typical methods. One is fundamental analysis. This method predicts the stock market by focusing on financial statements, interest rates, products, management, news and so on. This is used to get some insight on whether it is overvalued or undervalued. There are some studies to get some insight from news articles by using engineered method [3], [4], [5]. The other is technical analysis. This is a type of method which predicts the stock market by focusing on previous stock data, usually technical indicators such as moving average, relative strength index (RSI), stochastic oscillator and so on. There are many studies to predict the stock market by using such data and engineered approaches, such as Neural Network [6], Evolutionary Algorithms [7], Support Vector Machine [8], Neuro-Fuzzy [9], Hidden Markov model [10] and decision tree [11]. Concerning the other approaches, there are some studies that research the interrelationship between stock prices of the predicted stock and other time series data, such as foreign stock, temperature, audience rate and so on. Sung and So analyzed association rule for predicting changes in the Korea Compose Stock Price Index based on the time series data of various interrelated world stock market indices [12]. Johan, Huina and Xiaojun discovered that measurements of collective public mood states derived from large scale Twitter feeds correlate to the value of the Dow Jones Industrial Average over time. Moreover they also find an accuracy of 86.7% in predicting the daily up and down changes in the closing value by using its measurements [13]. In this approach, however, it is difficult to discover which data is interrelated with the predicted stock.

Therefore we propose a method that extracts interrelationships of changes in price between the predicted stock and various time series data, such as other stocks, world stock market indices, foreign exchanges and oil prices from real data automatically. We test our method for predicting the daily up and down changes in the closing value by using discovered interrelationships.

The rest of this paper is organized as follows: Section 2 discusses the proposed method. Experiments and results are reported in Section 3 while conclusions are the topic of Section 4.

II. Proposed Method

While there are some causes of changes in stock prices, information about other than the predicted stock should have effects on the predicted stock. For example, a stock related to exports is affected by foreign exchanges or foreign stocks. Therefore we aim to extract interrelations of changes in stock prices between the predicted stock and various time series data from real data, then predict the stock by using extracted interrelationships. This method is composed of two phases, interrelation discovery phase and prediction phase.

978-1-4673-1686-6/12/$26.00 ©2012 IEEE

A. Interrelation discovery phase

In this phase, we discover interrelationships between the predicted stock and various time series data. We investigate whether there is a specific fluctuation of referenced time series ahead of the rise or drop of the predicted stock. We define a specific fluctuation as a *variation pattern*, which is a representation how referenced time series data has changed. A variation pattern of referenced time series data is computed one-on-one with the predicted stock. Figure 1 shows the detailed description. How to compute the variation pattern is as follows.

1) Quantize referenced time series data: In this study, we quantize referenced time series data for simplicity of discovering interrelationships. How to quantize is as follows.

At first, referenced time series data is converted to rate of change of value:

$$C(t) = \frac{y(t) - y(t-1)}{y(t-1)}, \tag{1}$$

where $C(t)$ is the rate of change of value at time t, $y(t)$ is the value of referenced time series data at time t.

Then, $C(t)$ is quantized to five classes. We define five classes by comparing with previous change rates to maintain appearance frequency of each class. The expression (2) shows how to quantize.

$$Q(t) = \begin{cases} 0 & (C(t) > S_i, i = N \times 0.1) \\ 1 & (S_i \leq C(t) < S_j, i = N \times 0.1, j = N \times 0.3) \\ 2 & (S_i \leq C(t) < S_j, i = N \times 0.3, j = N \times 0.7) \\ 3 & (S_i \leq C(t) < S_j, i = N \times 0.7, j = N \times 0.9) \\ 4 & (otherwise) \end{cases} \tag{2}$$

where $Q(t)$ is the quantized number at time t, S_i is the ith $C(t)$ which is sorted in descending order over the past N days, in this paper N is 150.

2) Discover the variation pattern: A variation pattern represents how the referenced time series data has changed in the past ahead of the rise or drop of the predicted stock. Figure 2 shows an example of a variation pattern. The number at time t denotes the class number. In addition, # denotes not caring at the time. This pattern is searched by using Evolution Strategy (ES). ES searches the class number, whether caring or non-caring and the operator to extract good variation pattern. The operator denotes whether the class number cares a neighbor number. We define three operators to make the variation pattern flexible: A, B and C. The operator A denotes that the class number does not care a neighbor number, B denotes that the class number cares a large one, and C denotes that the class number cares a small one.

The variation pattern takes account of not only the previous data but the past M days data, 5 in this paper, as Fig. 2. ES tries to search the pattern whose hit rate and the number of prediction are high in learning period. The expression (3) and (4) show hit rate H and fitness function, respectively.

$$H = \frac{h}{n}, \tag{3}$$

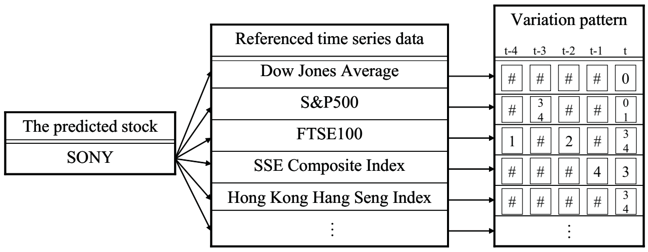

Fig. 1. Outline of Interrelation discovery phase.

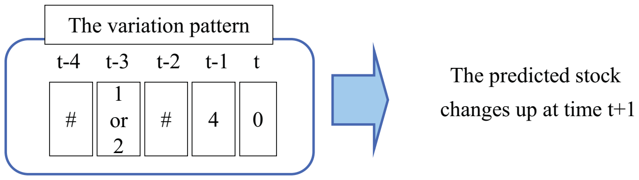

Fig. 2. Example of a variation pattern.

$$fitness = \begin{cases} H & (H < 0.7) \\ H + n & (otherwise) \end{cases} \tag{4}$$

where h denotes the number of correct predictions and n denotes the number of times that the variation pattern matches.

We discover two variation patterns with respect to each referenced time series data: the pattern in prediction of up changes and down changes.

B. Prediction phase

In the interrelation discovery phase, we obtain variation patterns with respect to each referenced time series data. It can not be said, however, that all of the obtained variation patterns are useful. Therefore we must make a choice as to which patterns are useful. In this study, we consider the pattern whose h is high is better. We set a threshold of h for discovering better patterns. The expression (5) shows the procedure of making the threshold th.

$$th = \overline{h} + 4\sigma, \tag{5}$$

$$\overline{h} = \frac{1}{N_{rd}} \sum_{i=1}^{N_{rd}} h_i, \tag{6}$$

$$\sigma = \sqrt{\frac{1}{N_{rd}} \sum_{i=1}^{N_{rd}} (h_i - \overline{h})^2}. \tag{7}$$

We consider the variation pattern whose h is higher than threshold is better. We define such referenced data as *likely interrelated data (LID)*. We predict the stock when one of LIDs matches with its variation pattern. If the number of LID is more than 10, in order to raise the reliability of the prediction we predict the stock when more than two of LIDs matches with its variation pattern.

TABLE I
PARAMETER SETTINGS FOR ES

Parameter	Value
Number of Generations	5,000
Generation alternation model	(1+4)ES
Mutation rate	1/(gene length)

TABLE II
PARAMETER SETS FOR THE VARIATION PATTERN

Parameter	Candidate value
Whether caring or not	True,False
Operator	A,B,C
Class number	0,1,2,3,4

III. EXPERIMENTS AND RESULT

We test our method on the stock market prediction. Table I and II show parameter settings for ES and candidates of parameters for the variation pattern at time t, respectively. The experimental settings are shown in Table III. We chose various referenced time series data and typical Japanese stocks for prediction.

Table IV and V show numerical evaluations in predicting the daily up and down changes in the closing value. In this area, it is better to predict the daily up or down in more than 55% hit rate. Atsalakis and Valavanis summarized 14 studies about stock market prediction [9]. These results were 57.1% hit rate on average and 68.3% in the best. Therefore this method can predict stock directions well, especially in the manufacturing industry. We think the reason for this is there seem to be more relationships in the manufacturing industry, such as buying or selling components from the company, capitalizing the company for technologies and so on. Moreover, our method is not only predicting the stock but also discovering likely interrelated data.

Figure 3 shows LIDs and variation patterns in predicting up change of the stock of Sony . All obtained LIDs are world stock market indices. Moreover all variation patterns express that index rises greatly just before the stock of Sony rises (at time t). It is often said that a stock related to exports is affected by foreign economies, and our method obtained such interrelations. The same interrelationships are obtained when we predict the stock of Hitachi, NEC, Nikon and Nissan, all of them are a stock related to exports.

Figure 4 shows LIDs and a example of a variation pattern in predicting up change of the stock of MITSUBISHI HEAVY INDUSTRIES (MHI). Almost all obtained LIDs are stocks which seem to interrelate with heavy industry, such as train manufacture, automobile component manufacture, steel manufacture and so on. Furthermore, MHI is the biggest stockholder of Mitsubishi Steel and has been involving in the business cooperation with The Kinki Sharyo. This is to say that our method discovered interrelationships which can be expressed why they have interrelationships.

Figure 5 shows LIDs and a example of a variation pattern in predicting down change of the stock of Mitsubishi Corporation (MC). Since MC is a wholesale firm, we can understand why the Dow Jones Average and S&P500 are discovered as LID. The others, however, seem not to interrelate with MC because there are no business alliance with MC or these stocks are not in the same kind of business area as MC. Nevertheless, they are useful in prediction because it goes well. Though we would not

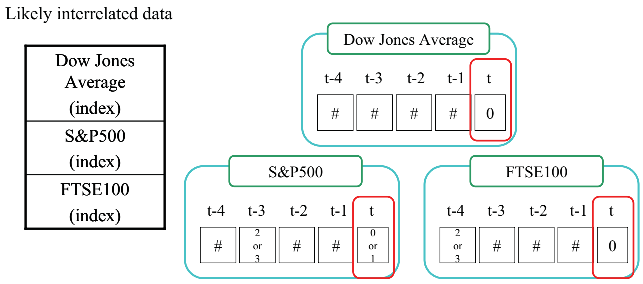

Fig. 3. Obtained variation patterns when we predict the stock of Sony.

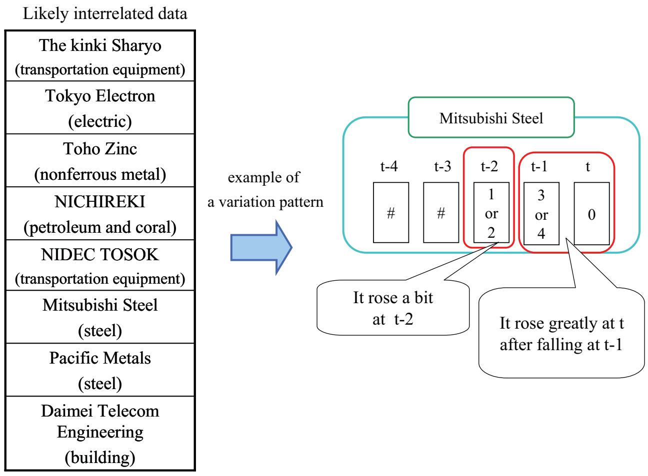

Fig. 4. Obtained variation patterns when we predict the stock of MITSUBISHI HEAVY INDUSTRIES.

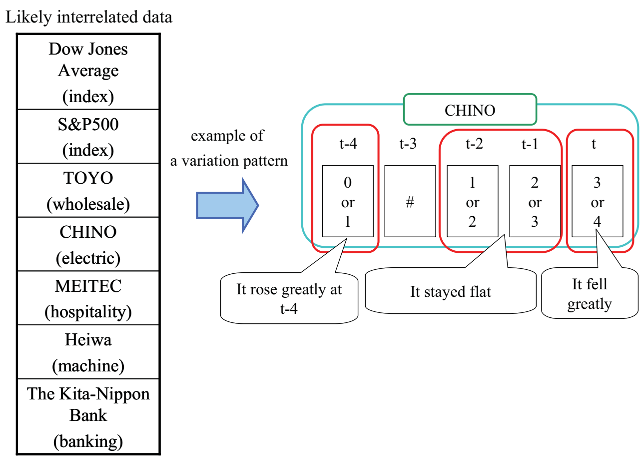

Fig. 5. Obtained variation patterns when we predict the stock of Mitsubishi Corporation.

explain why these interrelationships are discovered, we think it is also important to discover such interrelations because there may be some indirect interrelationships such as the butterfly effect. In addition, we have to analyze the reason why these are discovered because we may obtain new knowledge about the industrial structure.

IV. Conclusion

In this paper we proposed the methods that extract interrelationships of changes in prices between the predicted stock and various time series data, such as other stocks, world stock market indices, foreign exchanges and oil prices. Our method calculates variation patterns which represent how referenced time series has changed by using Evolution Strategy, then extracts likely interrelated data, and predicts the stock by using obtained interrelationships. We tested our method on the stock market prediction. Experimental results showed that our method can predict stock directions well, especially in the manufacturing industry. Obtained LIDs are not only stocks which is stockholder or has involved in the business cooperation with the predicted stock but also stocks which seem not to interrelated with the predicted stock.

In our future works, we will work on analyzing the interrelation between the predicted stock and other time series output as LID, and develop a trading strategy using our method.

References

[1] Alex A, Freitas, “Data Mining and Knowledge Discovery with Evolutionary Algorithms”, Springer, 2002.

[2] Ashish Ghosh, Lakhmi C. Jain, “Evolutionary Computation in Data Mining”, Springer, 2004.

[3] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Wai Lam, “Stock Prediction: Integrating Text Mining Approach using Real-Time News”, Computational Intelligence for Financial Engineering, pp. 395 - 402, 2003.

[4] Di WU, Gabriel Pui Cheong FUNG, Jeffrey Xu YU, Qi PAN, “Stock prediction: an event-driven approach based on bursty keywords”, FRONTIERS OF COMPUTER SCIENCE IN CHINA,Vol. 3, No. 2, pp. 145-157, 2009.

[5] Robert P.Schumaker, Hsinchun Chen, “A Discrete Stock Price Prediction Engine Based on Financial News”. Computer, Vol. 43, No. 1 ,pp. 51-56, 2010.

[6] Kyoung-jae Kim, “Artificial neural networks with evolutionary instance selection for financial forecasting” , Expert System with Applications Vol. 30, pp. 519-526, 2006.

[7] Pawel B. Myszkowski, Lukasz Rachwalski, “Trading rule discovery on Warsaw Stock Exchange using coevolutionary algorithms”, Proceedings of the International Multiconference on Computer Science and Information Technology, Vol. 4, pp. 81-88, 2009.

[8] Wei Huang, Yoshiteru Nakamori, Shou-Yang Wang, “Forecasting stock market movement direction with support vector machine”, Computers & Operations Research, Vol. 32, No. 10, pp. 2513-2522, 2005.

[9] George S.Atsalakis, Kimon P. Valavanis, “Forecasting stock market short-term trends using a neuro-fuzzy based methodology”, Expert System with Applications, Vol.36, pp. 10696-10707, 2009.

[10] Md. Rafiul Hassen, Baikunth Nath, Michael Kirley, “A fusion model of HMM, ANN and GA for stock market forecasting” , Expert Systems with applications, Vol. 33, No. 1, pp. 171-180, 2007.

[11] Jar-Kong Wang, Shu-Hui Chan, “Stock market trading rule discovery using two-layer bias decision tree”, Expert System with Applications, Vol.30, pp. 605-611, 2006.

[12] Sung Hoon Na, So Young Sohn, “Forecasting changes in Korea Composite Stock Price Index(KOSPI) using association rule”, Expert System with Applications, Vol.38, pp. 9046-9049, 2011.

[13] Johan Bollen, Huina Mao, Xiaojun Zeng, “Twitter mood predicts the stock market”, Journal of Computational Science, vol.2, pp. 1-8, 2011.

TABLE III
EXPERIMENTAL SETTING

Referenced time series data	1371 stocks listed on the Tokyo Stock Exchange, TOPIX, Dow Jones Average, S&P500, FTSE100, SSE Composite Index, Hong Kong Hang Seng Index, yen-dollar exchange rate, yen-GBP exchange rate, yen-Euro exchange rate, yen-AUD exchange rate, yen-NZD exchange rate, yen-CAD exchange rate, yen-CHF exchange rate, Euro-dollar exchange rate, WTI
Predicted stock (Industry)	Nippon Meat Packers, Inc. (food) COSMO OIL Co, Ltd. (petroleum and coral) ITOCHU Corporation. (wholesale) Mitsubishi Corporation. (wholesale) Hitachi, Ltd. (electric) NEC corporation. (electric) Sony Corporation. (electric) Nikon Corporation. (precision equipment) MITSUBISHI HEAVY INDUSTRIES, LTD. (machine) Nissan Motor Co, Ltd. (automobile) TOYOTA MOTOR CORPORATION. (automobile)
Training periods	4 January 1999 to 29 December 2006
Test periods	4 January 2007 to 30 December 2008

TABLE IV
NUMERICAL EVALUATIONS IN PREDICTING THE DAILY UP

		training period		test period	
Stock name	Number of LID	Number of prediction	Hit rate(%)	Number of prediction	Hit rate(%)
Nippon Meat Packers	8	519	68.4	125	52.8
COSMO OIL	8	437	68.9	117	52.1
ITOCHU Corporation	3	217	71.6	81	56.8
Mitsubishi Corporation	5	196	69.3	58	44.8
Hitachi	12	272	79.0	102	55.0
NEC	5	451	70.1	135	**66.7**
Sony	3	496	74.0	146	**68.5**
Nikon	12	302	80.1	74	**60.8**
MITSUBISHI HEAVY INDUSTRIES	8	449	69.3	98	59.2
Nissan Motor	11	246	80.1	103	**63.1**
TOYOTA MOTOR	2	187	70.1	50	**60.0**
Average			72.8		58.2

TABLE V
NUMERICAL EVALUATIONS IN PREDICTING THE DAILY DOWN

		training period		test period	
Stock name	Number of LID	Number of prediction	Hit rate(%)	Number of prediction	Hit rate(%)
Nippon Meat Packers	13	203	84.2	38	50.0
COSMO OIL	6	259	68.5	85	49.4
ITOCHU Corporation	8	368	69.2	127	**68.5**
Mitsubishi Corporation	7	358	69.1	152	**62.5**
Hitachi	10	238	75.6	88	**69.3**
NEC	16	672	76.2	175	**64.6**
Sony	8	693	68.1	190	59.5
Nikon	12	159	71.2	86	59.3
MITSUBISHI HEAVY INDUSTRIES	9	584	68.5	129	54.3
Nissan Motor	10	108	86.1	22	**63.6**
TOYOTA MOTOR	6	411	68.6	116	**65.5**
Average			73.2		60.6