Short-term Stock Price Analysis Based on Order Book Information

Kenichi Yoshida
Graduate School of Business Science, University of Tsukuba, Japan
yoshida@gssm.otsuka.tsukuba.ac.jp

Akito Sakurai
Graduate School of Science and Technology, Keio University, Japan
sakurai@ae.keio.ac.jp

keywords: efficient market hypothesis, high-frequency trading, time series analysis, stock market prediction, technical analysis

Summary

Efficient market hypothesis is widely accepted in financial market studies and entails the unpredictability of future stock prices. In this study, we show that a simple analysis can classify short-term stock price changes with an 82.9% accuracy. Our analysis uses the order book information of high-frequency trading. The volume of high-frequency trading, which is responsible for short-term stock price changes, is increasing dramatically; therefore, our study suggests the importance of analyzing short-term market fluctuations, an aspect that is not well studied in conventional market theories. The experimental results also suggest the importance of the new data representation and analysis methods we propose, neither of which have been thoroughly investigated in conventional financial studies.

1. Introduction

Efficient market hypothesis (EMH [Basu 77]) is widely accepted in financial market studies. It asserts that financial markets are “informationally efficient.” As a consequence, future stock prices change randomly. In other words, they are unpredictable [Malkiel 99].

In this study, we show that a simple analysis can predict short-term stock prices with an 82.9% accuracy. The volume of high-frequency trading (HFT [Chlistalla 11]) is increasing dramatically; therefore, an analysis of short-term stock returns is increasingly important. Our results suggest the importance of analyzing the short-term market by showing unexpected prediction accuracy.

Our analysis uses order book information from the Tokyo stock exchange (TSE, the third largest stock exchange in the world [Tokyo stock exchange 12]) from Jan. 2010 to May 2014. The total amount of data analyzed was approximately 6 terabytes. The volume of order book information clearly reflects the increase in HFT. Although these trading records are neither small nor easy to analyze, recent developments in high-performance computing (i.e., “big data” processing), enables an analysis of this magnitude and complexity to be performed.

During our study, we attempted to use common techniques and data representations used in most conventional financial studies. However, the results were not pleasing. This indicated that we would have to extract information from UDP packets by designing new data representation methods. The poor prediction ability of conventional techniques suggests the importance of a new approach toward data representation and analysis techniques.

Chapter 2 of this paper first surveys and analyzes related work to determine the limitations of existing approaches to clarify the motivation of this research. Chapter 3 explains the proposed data representation and the analytical method. Chapter 4 reports on the experimental results. Chapter 5 discusses related topics. Finally, Chapter 6 concludes our findings.

2. Related Works

EMH asserts that future stock prices are difficult to predict. However, some researchers have challenged this statement (e.g., [Huang 94, Gallagher 02, Chordia 04, Gilbert 10, Gay Jr 11, Hsu 10, Menkhoff 10, Deng 15]). For example, [Huang 94] and [Chordia 04] analyzed relationship between individual stock returns and order imbalance. [Huang 94] extracted prediction rules, and [Chordia 04] made regression model. However, none of them reported the prediction model for HFT with comparative accuracy reported in this paper. The recent use of social
network services (SNSs, e.g., Twitter) data also has been studied (e.g., [Bollen 11, Sprenger 13, Sul 14, Deng 14]). Studies based on SNS data seem to analyze the occurrence of information asymmetry. These studies analyze the information propagation process through SNSs and found situations in which information asymmetry exists. Information symmetry is an important assumption of EMH; therefore, a prediction ability in such a situation is not surprising.

In our study, we aim to find situations in which information asymmetry exists in HFT. The order book is the source of information from which we try to determine the asymmetry. The order book contains information about ask and bid orders for each stock at a certain price. Each order represents the price that each trader recognizes as being adequate. In HFT, each trader experiences a different time lag when processing information. Even with high-speed computers and with a high-speed network connection, such asynchrony is inevitable. With an adequate data representation, we try to analyze the effects of such asynchrony.

As the amount of HFT is increasing, there has been a proliferation of primers and introductory handbooks on this topic (such as [Lewis 14] and [Patterson 13]). However, there still exists room for a comprehensive study of the characteristics of HFT. For example, to the best our knowledge, there is no simple report on the predictability of stock returns through HFT. EMH might discourage such studies. Current studies analyze different aspects, such as liquidity that formed the subject of [Beltran-Lopez 12] and [Danielsson 12].

In conventional studies of the financial market, methods from statistics and signal processing, such as the autoregressive (AR) model [Champernowne 48], the auto-regressive moving average (ARMA) model [Box 13], and the auto-regressive integrated moving average (ARIMA) model [Box 13] are used. These methods model time series data with a simple combination of elementary functions. For example, the AR model represents a sequence of events over time as \( y_t = \phi y_{t-1} + w_t \). This equation relates the value assumed by \( y \) at time \( t \) to another variable \( w_t \) and to the value \( y \) at the previous time. Here, \( t \) represents a moment in time based on a fixed time interval. Our experimental results show the inappropriateness of these conventional approaches, and suggest the importance of the proposed new data representation and analysis methods.

3. Order Book and its Representation

3.1 Order book in Arrownet

The TSE has been distributing real-time stock exchange information through a high-speed trading network named “Arrownet” since Jan. 2010 [Tokyo stock exchange 12]. When the TSE processes an order, it sends a UDP packet that contains information about the order, i.e., the stock ID, size of order, and ask/bid price. In particular, each UDP packet contains the eight lowest ask orders and eight highest bid orders (See Figure 1). The packet can therefore be described as a list of orders that records the interest of buyers and sellers on a particular stock, and is essentially an order book for the stock (see Table 1 for example). A comparison of the current packet with its predecessor for the same stock enables the extraction of the information of the corresponding order.

Figure 2 compares the data representations of orders. As explained before, each UDP packet corresponds to a single order. Figure 2 (1) shows the actual order events in real world trading. In the real world, each order event occurs at a random interval. On the contrary, conventional methods use fixed intervals, as shown in Figure 2 (2), which shows a typical time series representation of the same order events. Because a fixed interval is used to represent data, the timing information of events is lost in this representation.
Short-term Stock Price Analysis Based on Order Book Information

Although the conventional time series representation enables mathematical analyses of quoted price changes, its use of a fixed time interval disregards the timing information of events. In addition, in most cases, it does not distinguish between ask and bid orders, but only represents the quoted price of the stock. In the example shown in Figure 2 (1), the interval between events A2 and B1 is short, but the interval between events B1 and A3 is much longer. Such information is lost in the example shown in Figure 2 (2). Furthermore, ask and bid orders are also not distinguished in the example in Figure 2 (2).

To represent the information that is disregarded in conventional time series representations, we have designed two event-based representations (Figure 2 (3) and (4)). Event-based representation 1 (Figure 2 (3)) represents the order book information at the moment when each order event is processed. In particular, it represents the stock ID of the event, the most recently quoted price: ask/bid size/price. Other information included in our design to improve the analysis is also stored (See Section 4 · 1f o r details). Event-based representation 2 (Figure 2 (4)) represents the same information as event-based representation 1, but only the one at the time at which a trade of the stock takes place.

3.2 Preliminary experiments

We conducted several preliminary experiments which suggested that the conventional representation of time se-

Fig. 2

Order Book Information
Time | Quoted Price | Number of Asks | Number of Bids

Fig. 2 Representation of order events

Fig. 3 Correlation coefficients between $AV E_{am}$ and returns

AV $E_{am}$ simply sums up the weighted difference between ask/bid levels and the most recent quoted price of the stock at the time when an order is processed. We tried to extract the interest of buyer and seller on the quoted price by summing up $AV E_{t}$ stored in each UDP packet, i.e., each order. Thereby, the price of the eight lowest ask orders (i.e., ask price levels) and the eight highest bid orders (16 orders in total) are extracted from each UDP packet.

Figure 3 shows the Spearman’s rank correlation coefficients between $AV E_{am}$ and the corresponding stock returns. The coefficients were calculated by first selecting the 100 most frequently traded stocks during the period from 2010 to 2012 (3 years or 680 days). Then, we calculated the difference ($d_{am}$) between the first quoted price of the morning period and the last quoted price of the morning period for each stock. Finally, we calculated Spearman’s rank correlation coefficient between $AV E_{am}$ and $d_{am}$ each day. “AVE vs. AM” in the figure shows the results. Here, the $y$-axis shows Spearman’s rank correlation coefficients. The coefficients are sorted according to the values. As shown in the figure, the data is weakly correlated.

However, this weak correlation disappears when we cal-
culate the coefficient between $AVE_{am}$ and the stock returns for the afternoon period (i.e., the difference between the last quoted price of the morning period and the last quoted price of the day). The plot “AVE vs. PM” in Figure 3 shows the coefficients.

Although the non-existence of a correlation between $AVE_{am}$ and the afternoon returns suggests the impossibility to predict returns even in half a day, the existence of a weak correlation between $AVE_{am}$ and the morning returns suggests the possibility that short-term returns may be predicted. If $AVE_{am}$ did not have any relation with stock returns, the difference between “$AVE_{am}$ vs. AM” and “$AVE_{am}$ vs. PM” would lack reason for its occurrence. The observed difference seems to suggest the existence of a short-term correlation.

The next step in our study was to determine whether the weak correlations are utilized by the conventional representation, i.e., fixed interval representation of quoted prices. We accomplished this by building the following linear regression model learned by support vector regression (SVR):

$$\text{Quoted Price}_{t+10} = w^T \Phi(x_t) + b + \epsilon_t$$

Here $x_t$ is the vector of Quoted Price, and $AVE_t$ ($i = t, t-1, \ldots, t-30$). The interval is one minute. Figure 4 shows the squared correlation coefficients of the true and regressed Quoted Price, which are sorted in increasing order. “With AVE” corresponds to the regression with $AVE_t$s in explanatory variables and “Without AVE” corresponds to the regression without $AVE_t$s.

Although the weak correlations found in Figure 3 suggest the existence of a short-term correlation, Figure 4 shows that the SVR model failed to reproduce it.

We summarize the results of the preliminary experiments as follows. When the SVR model was trained to predict future quoted prices using the time series data from the previous 30 min by using the representation of conventional time series analysis, the SVR model was unable to confirm the existence of a weak short-term relationship, even though the correlation analysis shown in Figure 3 suggests that such a relationship exists. This led us to conclude that the conventional time series data representation is responsible for causing the SVR to perform poorly by removing important information from the data. The event-based representation shown in Figure 2 was inspired by this observation.

### 4. Experimental Results using Event-based Representation

#### 4.1 Event-based representation and Attributes

This chapter reports the results of experiments that use the TSE order book information from Jan. 2010 to May 2014. We analyzed the UDP packet information that was exchanged through the high-speed trading network named “Arrownet” [Tokyo stock exchange 12]. The total amount of data that was analyzed was approximately 6 terabytes.

Event-based representations (Figure 2 (3) and (4)) were used in the analysis.

The analysis used libSVM [Chang 11] and J48 (an implementation of C4.5 in weka [Hall 09]). The default parameters were used in the experiments to solve the three class problem (“up,” “down,” and “unchanged”). The attributes that were used are:

**Average price in orders:** $e_1$

As explained before, $AVE_t$ is the difference between the most recent quoted price and the average price of most recent orders. $e_1$ is the sum of $AVE_t$s between the time of the most recent trade and the time of the current order.

**Number of asks and bids:** $e_2, e_3, e_4$

Number of UDP packets for “ask ($e_2$)” and “bid ($e_3$)” after the most recent trade. Here, $e_4 = e_2 - e_3$.

**Number of lower asks and higher bids:** $e_5, e_6, e_7$

Number of UDP packets that ask with price lower than the most recent quoted price ($e_5$), and number of UDP packets that bid with price higher than the most recent quoted price ($e_6$). $e_7 = e_5 - e_6$.

**Number of lower bids and higher asks:** $e_8, e_9, e_{10}$

Number of UDP packets that ask with price higher than the most recent quoted price ($e_8$), and number of
UDP packets that bid with price lower than the most recent quoted price \( e_9 \). \( e_{10} = e_8 - e_9 \).

**Number of asks and bids with quoted price:** \( e_{11}, e_{12}, e_{13} \)

Number of UDP packets for ask and bid with the same price as the most recent quoted price after the corresponding trade \( e_{11}, e_{12} \). \( e_{13} = e_{11} - e_{12} \).

**Differences between the recent quoted price and its predecessor:** \( e_{14} \)

This may either be the latest difference or a series of the values up to the four most recent differences.

As explained above, \( e_1 \) is designed to capture the average of peoples’ perception of the stock price. The variables \( e_2, e_3, ..., e_{13} \) are designed to count the number of people who wish to ask and bid and \( e_{14} \) is designed to express the trend of the stock returns.

### 4.2 Classification accuracy based on order book information

Figure 5 shows the average classification accuracy of the direction of stock price changes based on the order book information. These results are based on data from the TSE from Jan. 2010 to May 2014. The 100 most frequently traded stocks are used in the experiments. The order book history for each stock is recorded at the time when an order is processed. The values of quoted price, ask/bid size and price are extracted from the data to make the order history of each stock. Then subsequences of order history, whose length are 10,000, are extracted to make training sets. They are represented with event-based representations 1 and 2. Using each training set, the SVM and J48 models are trained to learn a classification model, which solves a three class (i.e., price will go “up,” “down,” or remain “unchanged”) problem. For each training set, the subsequent 1000 orders are extracted to construct the corresponding test set. Figure 5 shows the average prediction accuracy.

ZeroR classifies all the data according to the mode of the target variable in the training data. Because a few samples are “unchanged” and about half of the samples are either “up” or “down,” the classification accuracy of ZeroR is approximately 50% as shown in Figure 5.

\( e_{14} \) in Figure 5 shows the average classification accuracy (58.7%) by J48 which uses only \( e_{14} \) as the explanatory variable. We conducted the experiments by varying the length of \( e_{14} \) to determine whether the classification accuracy would be affected by the length of the quoted price history. However, we found the prediction accuracy did not differ much. Thus, we only used the most recent \( e_{14} \) values as the data for the quoted price history.

“SVM_R2” and “J48_R2” in Figure 5 show the average classification accuracy of SVM and J48 based on event-based representation 2 with variables \( e_1, e_2, ..., e_{13} \). The result of “J48_R2” (i.e., 79.3%) clearly shows the classification ability of the order book information. Note that “SVM_R2” and “J48_R2” only use the most recent value for \( e_1, e_2, ..., e_{13} \) and do not depend on its past history \( e_{14} \).

The high classification accuracy of 79%, shown in Figure 5, requires careful explanation, because this result does not suggest that there is a prediction method capable of achieving a 79% prediction accuracy. The classification was performed based on information extracted from the latest UDP packet and possibly previous UDP packets. The output of the learned classifier is the direction of possible quoted price changes by the order, i.e., the direction of the quoted price changes if the order results in a trade, or, otherwise, the direction of the quoted price changes in future. In other words, the change has already been realized once a trade has taken place; therefore, the output is just a classification but not a prediction.

Although we will show the true prediction accuracy which can be used in the actual trading shortly, the important finding in Figure 5 is “Short-term stock return does not walk randomly. The order book can classify stock return movement.” We can use Arrownet to classify stock return movements in HFT. Thus, the famous assertion “informational efficiency of financial markets” does not exist in HFT. The time lag existing in computers and networks, which is inescapable in HFT, may be a cause of such inefficiency.

“SVM_R1” and “J48_R1” show the average prediction accuracy based on \( e_1, e_2, ..., e_{13} \). These results are based on event-based representation 1. Under the representation, the expected output, or the target output of each predic-
tion, is the direction of change of the next quoted price. Therefore, this setting is close to actual trading. Though the prediction accuracy is only relatively high (60.7% by SVM and 61.9% by J48), it is higher than the result based on the history of quoted prices (i.e., 58.7% of $e_{14}$ in Figure 5).

Although the results of the preliminary experimentation shown in Figure 3 only suggest the prediction ability of order book information, the results in Figure 5 clearly confirms this. We believe the latter is mainly caused by the data representation we used. The results shown in Figure 4 exhibit inability of prediction and are based on the traditional time series representation while the results shown in Figure 5 suggest the prediction capability of our approach and are based on event-based representation.

Note that the results obtained with J48 are a slight improvement over those obtained with SVM. SVM models data by using a smooth function of explanatory variables. Although the training method for parameters of the function is different, traditional methods of time series analysis such as ARIMA also models data by applying a smooth function. However, J48 achieved better results using a totally different framework, i.e., a classification tree. These results indicate the possible inappropriateness of traditional time series analyzing methods (i.e., the use of fixed interval time series representation and smooth functions) for the analysis of HFT. See Chapter 5 for further discussion.

4.3 Improvement of accuracy

Because the prediction accuracy is important for actual stock trading, we have attempted to improve this accuracy and the results are reported in this chapter. The promising results we achieved by using event-based representation and J48, prompted us to use these two methods as the base framework, to aim to improve the prediction accuracy with additional means.

Figure 6 shows the classification/prediction accuracy of J48 using both historical data of the quoted price $e_{14}$ and order book information $e_1, e_2, ..., e_{13}$. The accuracy that was achieved when event-based representations 1 and 2 (“With1” and “With2” in the figure) were used with $e_{14}$. On the other hand, “Without1” and “Without2” show the accuracy without the use of $e_{14}$. From this, we concluded that the use of historical data together with order book information leads to a slight improvement of accuracy compared to when the order book information is excluded.

The improved accuracy we achieved with event-based representation 2, led us to perform additional experiments to determine the prediction accuracy of J48 model, which is trained with event-based representation 2, but is tested on the test set from event-based representation 1. With this setting, the target label is the quoted price change direction that is realized by the training data. This means that the output is just a classification and not a prediction. However, when the model is applied to the test data with event-based representation 1, because the input, i.e., an order, will not necessarily be a fulfilled order, the output label could be interpreted as a prediction of the direction of the next quoted price change. Based on this interpretation, we could use the model in real trading and calculate the expected prediction accuracy for the situation.

The prediction accuracy of a model trained with event-based representation 2 and tested with event-based representation 1 is 66.5% as shown in Figure 7 next to the label “Training.” This value is in between the values of event-based representations 1 (Rep.1 in the figure) and 2 (Rep.2 in the figure, Figure 7 Rep.1: 62.9% and Rep.2: 81.1% respectively*1). Although we have not yet identified the cause of this improvement (i.e., the difference in accuracy between “Rep.1” and “Training”), this phenomenon is interesting, and requires further study.

The last experiment we conducted as part of this study was to examine the number of data sets that would be required for training. During the experiments, we noticed fluctuations in the prediction accuracy. Figure 8 shows the distribution of the prediction accuracy for a single stock.

*1 Figure 7 Rep.1 shows the same value that was shown in Figure 6 (With1) for comparison purposes.
Short-term Stock Price Analysis Based on Order Book Information

The number of orders for this stock between Jan. 2010 and May 2014 was approximately 77,000,000, and we extracted 7,700 training sets, each of which consisted of 10,000 orders. Using the subsequent 1,000 orders of this stock, we calculated the prediction accuracy. Figure 8 plots each accuracy and the results vary greatly as can be seen in the figure. As we suspected this variation to be caused by the use of a training set of insufficient size, we performed the same experiment again after increasing its size.

Figure 9 shows the results of experiments with a larger size training set. In this experiment, five days of order histories were used as the training data set, and orders placed on the next single day were used for the test data set. As shown in Figure 9, the classification accuracy with event-based representation 1 increased from 81.1% to 82.9%, whereas the prediction accuracy with event-based representation 2 increased from 66.5% to 67.7% and the distribution of accuracy becomes narrower (See Figure 10).

Figure 11 compares the prediction accuracy of the models trained with one of the attributes $e_1, e_4, e_7, e_{10}, e_{13}$ and $e_{14}$. These results compare the explanatory power of order book information. As shown in Figure 11, the most informative attribute is $e_{13}$ other than $e_{14}$. Figure 12 shows an example of the learned rule during the experiments with attributes $e_{13}$ and $e_{14}$. As shown in Figure 12, $e_{13}$ (i.e., the number of asks and bids with the most recent quoted price) is sufficient to classify the future stock price change direction with accuracy 82.9%.

It is difficult to understand why this simple rule was not reported before. However, as far as the short-term prediction of HFT is concerned, it is clear that stock returns do not show a random walk. The institutional investors who have direct access to the trading system may have an unfair advantage over individual investors who do not...

\footnote{\textsuperscript{2} The size of the training/test data sets differed in terms of the stocks and date of trades. The average size of the training/test data sets for the data shown in Figure 9 were approximately 60,000 and 6,000 respectively.}

5. Discussion
have such high speed access. Although we did not investigate this phenomenon in other HFT markets, such as retail foreign exchange trading, the simplicity of the found rule suggests the generality of our finding. However, this issue would have to be investigated in future.

Figure 13 shows another possibility of unfairness. We conducted an experiment on the prediction accuracy of learned rules. The prediction accuracy of the learned rule decreases day by day. However, as Figure 13 shows, the speed of degradation is slow. This stability of the learned rule facilitates implementation of the system, because investors who wish to use the trading system with the found rule can replace the rules every weekend, as J48 requires less than an hour of learning time. Investors can update the system during weekends.

The last topics we have to discuss in this study are the comparison of J48 and SVM. We have conducted various experiments in which J48 tends to outperform SVM. However, the simplicity of the hidden rule behind the stock price changes, as shown in Figure 12, does not explain the inability of SVM to perform better than J48.

Figure 14 shows examples of the decision boundary found by J48 and SVM. As shown in Figure 14, the boundaries of SVM and that of J48 appear similar. The capability of using tilted boundaries (See Figure 14 (2)) or smooth nonlinear boundaries (Figure 14 (3)) would suggest that SVM has an advantage. However, the results suggest that the opposite is true.

---

Fig. 12 An example of learned rules

Fig. 13 Stability of the learned rule

Fig. 14 Boundary of J48 and SVM
Our assumption for the reason for this phenomenon is the difference between the penalty functions that J48 and SVM use for learning. The soft-margin formulation of SVM uses the distance \(|t - (w^T \phi(x) + b)|\) as the penalty for each miss-classified data. The penalty used by J48 is simpler, i.e., entropy or information gain. The evaluation measure for our experiments, i.e., the classification/prediction accuracy is the ratio of the number of samples that are correctly classified/predicted versus those that are not. The ratio is not dependent on how far from the boundary the miss-classified samples reside. Therefore, the penalty function used by J48 seems to be more adequate than that of SVM to express the errors for the prediction in the change of the stock price. Because a similar distance, such as \((t - (w^T \phi(x) + b))^2\), is used in most AR models and their descendants, they may suffer from a similar problem.

6. Conclusion

In this study, we analyzed order book information of the TSE (Tokyo Stock Exchange, the third largest stock exchange in the world) from Jan. 2010 to May 2014, and showed that a simple rule is capable of classifying the change in direction of short-term stock prices with an 82.9% accuracy and of predicting it with an accuracy of 67.7% by using the order book information of HFT (High-Frequency Trading).

As the volume of HFT, which forms short-term stock exchanges, is increasing dramatically, our study suggests the importance of short-time market analyses that are not well studied in conventional financial market theories.

The experimental results also suggest the importance of new data representation, i.e., event-based representation, and the analysis methods that are not well studied in conventional financial studies. The use of the penalty function to determine optimal model parameters is important to identify the model with optimal performance.

Although the prediction accuracy is less than the classification accuracy, it is still 67.7% and can be applied to actual trading. This level of accuracy may suggest that institutional investors who can access the trading system directly have an unfair advantage over individual investors who do not have such high-speed access.

Acknowledgments

This work was supported in part by JSPS KAKENHI Grant Number 25280114 and 25330266.

References


[Sprenger 13] Sprenger, T. O., Tumasjan, A., Sandner, P. G., and


Received July 27, 2015.

Author’s Profile

Kenichi Yoshida (Member)
received the Ph.D. in engineering from Osaka University in 1992. In 1980, he joined Hitachi Ltd. Since 2002, he has been a Professor at University of Tsukuba, Tokyo, Japan. His current research interest includes application of Internet and application of machine learning techniques.

Akito Sakurai (Member)
graduated from the University of Tokyo in 1975, completed his master’s course in 1977, and joined Hitachi, Ltd. In 1978 he earned an M.S. degree in Computer Science at the University of Illinois. In 1989 he moved the Advanced Research Laboratory, Hitachi, and in 1996 the Central Research Laboratory. In 1998 he became a professor at the Japan Advanced Institute of Science and Technology. He has been a professor at Keio University since 2002. He holds D.Eng. degree (The University of Tokyo, 1993).