Multiple Kernel Learning on Time Series Data and Social Networks for Stock Price Prediction

Shangkun DENG, Takashi MITSUBUCHI, Kei SHIODA, Tatsuro SHIMADA, Akito SAKURAI
Faculty of Science and Technology, Keio University
3-14-1 Hiyoshi, Yokohama-shi 223-8522, Japan
dsk8672@gmail.com , {mitsubuchi, shioda, shimada, sakurai}@ae.keio.ac.jp

Abstract

This paper proposes a stock price prediction model, which extracts features from time series data, news, and comments on the news, for prediction of stock price and evaluates its performance. In this research, we do not take account of text contents of news and user comments, but just consider numerical features of news and communication dynamics appeared in comments on the Web as well as historical time series data. We model the stock price movements as a function of these input features and solve it as a regression problem in a Multiple Kernel Learning regression framework. Experimental results show that our proposed method consistently outperforms other baseline methods in terms of magnitude prediction measures such as MAE, MAPE and RMSE for three companies' stocks. They specifically show that the features other than stock prices themselves improved the performance.

Keywords

Stock Price Prediction; Human Factors; Multiple Kernel Learning; Time Series Data; Social Networks; Communication Dynamics

I. Introduction

Lots of studies have been done to estimate and predict the movement of stock prices by using statistics and other forecasting methods based on the historical stock price or volume data and Technical Analysis. From previous works [1, 2], we find that technical analysis is one possible way to accomplish successful prediction in the Stock and Foreign Exchange (FX) market.

In addition, during the last decade, lots of researchers used machine learning technology such as Genetic Algorithm (GA), Neural Networks (NNs), or Support Vector Machine (SVM) to predict prices in the stock or FX market; e.g., D. Fuente et al. [3] applied GA to the trading of stock markets, F. Allen & R. Karjalainen [4] applied GA to generating trading rules. T.H. Hann & E. Steurer [5], and A.-S. Chen & M.T. Leung [6] applied neural networks to the financial prediction and found that they outperformed linear models. J.T.-Y. Kwok [7] and L.J. Cao [8] applied SVM to the prediction of stock market and obtained good performances.

However, human factors have significant impact on the movement of stock prices. To quantify the impacts we could turn to the Internet. In the Internet and information Age, numerous people post their comments on the website. Therefore, analyzing communication dynamics on the Internet and using the results to mine interesting correlations with stock price movements may provide some new insights into relations between stock prices and communication patterns. Previous works such as in [9] modeled communication dynamics and found the correlation with external events. Previous works [10, 11] also model communication dynamics in social networks. Munmun De Choudhury et al. [12] defined some kinds of communication dynamics in user comments and used a SVR framework to learn and to predict stock prices. Good performances in their testing period on certain stocks show that mining communication dynamics is an alternative way for stock prices prediction.

In recent years, many researchers have used a method to deal with the problem of selecting suitable kernels for different feature sets. This method is called Multiple Kernel Learning (MKL) [13]. A strong point of MKL is that it allows us to combine different kernels when it is better to use different kernels for different input features. MKL mitigates the risk of erroneous kernel selection to some degree by taking a set of kernels and deriving a weight for each kernel such that predictions are made based on the weighted sum of the kernels. Some researchers have applied MKL to prediction in the FX or stock market; Fletcher, Hussain & Shawe-Taylor [14] applied MKL to the limit order book for predicting and trading on the FX market; Luss & d'Aspremont [15] applied MKL for predicting abnormal returns from news using text classification; Shie-Jue Lee et al. [16] applied MKL to the prediction of prices on the Taiwan stock market, although the results looked a little better than some conventional methods, the prediction also looked like stating the current value; Deng et al. [17] applied MKL on the prediction and trading on FX rate and obtained good results.

In our paper, we extract features not only from time series data (price and volume) source, but also from social networks source (in this paper, we investigate on Engadget). We use a multiple kernel learning framework to learn and predict the stock prices for three Internet Technology companies: Microsoft, Amazon, and Google. The analysis of communication dynamics in this research are mainly based on [12] but with some changes to make them suit our purpose.

Our contribution here is twofold. First, in our research, we extract features from both time series data source and social

☆This research is partially supported by Global-COE program of Keio University

978-0-7695-4607-0/11 $26.00 © 2011 IEEE
DOI 10.1109/ICMLA.2011.99

network source, in contrast to, e.g., [12] where they considered only communication dynamics to predict stock price movements. In [15], they used both news text data and time series data as input source for prediction of abnormal returns and their proposed model obtained better results than that of baseline methods which considers only one source of them. It gives us an inspiration to have a trial on the prediction of stock prices by extracting features from both time series data and social networks in Internet. [15] focuses on text mining of news with time series data but our research focus on mining from communication dynamics of news and comments with time series data.

Second, we use a MKL model to optimally combine the features of stock prices, stock volume, numerical features of news and communication dynamics of comments, in contrast to, e.g., [12] where they use only single kernel for SVR. From previous works [5, 6, 7, 8], we know that Neural Networks and Support Vector Regression are good models to predict stock prices, especially SVR. However, if we use the features of both time series and communication dynamics as the input, because they could have different properties, we should consider using different kernels for the input features from different kind of sources. But clearly it is not easy to find good kernels manually. Therefore, we use MKL to solve this problem. In addition, the coefficient results we will obtain in the MKL training step will show us relative importance of different sources on the prediction of stock price movements.

In our experiments, we use a Window-Shift method for training and testing because the dynamics of stock market would change and we try to make our model adapt to the changes in stock market. We supposed that the dynamics of stock market are changing slowly but could be considered relatively the same in the training period. This property is not supported by some other researches or evidence and, in fact, the money market sometimes changes abruptly. However, our experimental results showing that the proposed system obtains stable and better results in magnitude prediction than other baseline methods demonstrate that the assumption is not absurd.

The rest of this paper is arranged as follows. Sections II simply describes the SVM regression, multiple kernel learning, and communication dynamics which are mainly based on previous work [12]. Section III describes our proposed method and each component of the proposed model. Section IV details our experimental design. Section V shows our experimental findings and discusses their impact in terms of MKL coefficients of each source, i.e., time series, news frequency, and communication dynamics of comments on stock price prediction. Section VI concludes the paper and provides some future research directions.

II. SVR, MKL, and Communication Dynamics

A. Support Vector Regression

Support Vector Regression (SVR) is a version of Support Vector Machines (SVM) [18] with some distinct advantages. For example, SVR solves a risk minimization problem by balancing the empirical error and a regularization term, where the risk is measured by Vapnik's ε-insensitive loss function. In addition, SVR usually estimates a set of linear functions defined in a high dimensional feature space. Furthermore, SVR is well-known for its ability to perform well when there are many relevant features.

The regression function can be estimated by minimizing a regularized risk function

$$\min \ \frac{1}{2}\|w\| + C\frac{1}{l}\sum_{i=i}^{l} L_\varepsilon \qquad (1)$$

$$L_\varepsilon = \begin{cases} |y_i - w * \Phi(x_i) - b| - \varepsilon, & \text{if } |y_i - w * \Phi(x_i) - b| \geq \varepsilon \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

where w is a weight vector to be defined to determine the maximum margin hyper-plane.

B. Multiple Kernel Learning

A normal SVM is applied to a single kind of features. In our experiments, we use the Multiple Kernel Learning (MKL) to integrate features of different kinds. With MKL, we train an SVM with an adaptively-weighted combination of kernels which fuses different kinds of features. The combined kernel is as follows:

$$K_{comb}(x,y) = \sum_{j=1}^{K} \beta_j K_j(x,y) \qquad (3)$$

$$\text{with } \beta_j \geq 0, \quad \sum_{j=1}^{K} \beta_j = 1 \qquad (4)$$

where β_j are weights to combine sub-kernels $K_j(x,y)$. MKL will estimate optimal weights from training data. By preparing one sub-kernel for each feature set and estimating weights by MKL, we obtain an optimal combined kernel.

Sonnenburg et al. [19] proposed an efficient algorithm of MKL regression to estimate optimal weights and SVM parameters simultaneously by iterating training steps of a normal SVM. In our experiment, we use the MKL library included in the SHOGUN toolbox.

C. Communication Dynamics

TABLE 1 A list of features based on news and communication dynamics

No	Features based on news and communication dynamics
1	Frequency of News
2	Frequency of Comments
3	Average and standard deviation of response time of comments
4	Average and standard deviation of comments length
5	Number of Loyals and Outliers (include "Guests")
6	Number of Rank of comments
7	Number of Early and Late Responder

We now introduce some communication dynamics of our dataset on the Engadget website. This social network is characterized by two kinds of data source: news and comments. In addition, every comment on news is also marked for its significance by other users. There are five levels of significance of comments: highest ranked, highly ranked, neutral, low ranked, lowest ranked. The level associated with a comment at any instant of time represents the composite significance indicated by all users. Table 1 shows a list of these communication dynamics used as features. These features and definitions are based on [14] but we also make some changes in No.5 and No.7 in the list. In No.5, we also consider "Guests" as "Outliers", and in No.7, we set two different parameters for Early and Late responder respectively.

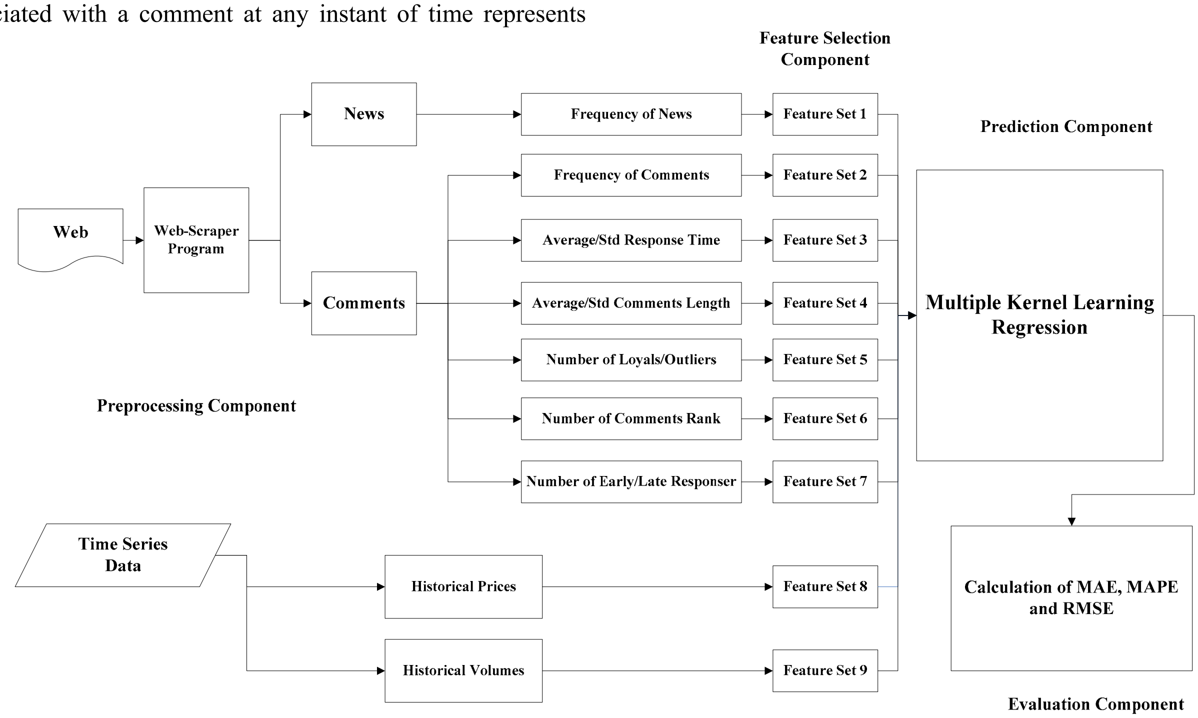

Figure 1. Diagram of Proposed Model

III. Proposed Model

Fig. 1 shows an overview of our proposed model. Our proposed model is composed of four components :

- Raw Data Preprocessing Component
- Feature Selection Component
- Stock Prediction Component
- Performance Evaluation Component

A. Raw Data Preprocessing Component (RDP component)

The Raw Data Preprocessing (RDP) Component processes the raw data to be used for our experiments. For the web source, RDP Component downloads the news and comments with their attributes such as publishing time, level of comments, text of news and comments. For the time series source, the RDP component downloads the historical price and volume data from Google Finance.

B. Feature Selection Component (FS component)

The Feature Selection (FS) component extracts the features we need from the data we download in RDP component. For the time series source, FS Component extracts the historical closing price and transaction volume every day for three companies, and for web sources, it extracts numerical features of the news and communication dynamics of comments that we defined in Section II.C from news and comments we downloaded in RDP component. Before the process of FS component, we should first set the values for each parameter in this component.

C. Stock Prediction Component (SP component)

After FS component, we use a Multiple Kernel Learning regression to learn and then to predict the change rate and stock price for the next trading day. In MKL model, we use one linear kernel and one Gaussian kernel for each input feature set, and we simply set the default values for the parameters of the Gaussian kernel.

D. Performance Evaluation Component (PE component)

In the performance evaluation component (PE component), we use MAE, MAPE, and RMSE as the evaluation measures to evaluate the goodness of proposed and baseline models.

IV. Experimental Design

A. Data Sources

Our data for training and testing are from two sources: Google Finance and Engadget. The Time Series Data of the daily stock price and volumes are obtained from Google Finance Website [20], and we downloaded the news and comments data from Engadget [21].

In our experiment, we select three companies for our experiment: Amazon, Microsoft and Google, which are three important companies in US stock markets, since Engadget is a special community for Internet and Information Technology.

Table 2 shows the number of news and comments for three companies from January 1, 2006 to August 15, 2008.

TABLE 2 Number of news and comments

Company	Number of News	Number of Comments
Amazon	298	9253
Goolge	555	17594
Mircrosoft	1803	69023

B. Input Feature Sets and Output

We use a preceding one week (7-days) data as the input features to predict stock prices of a day, that is, we assume that the stock movement of a company on a certain weekday depends on the stock prices and communication activity in the preceding 7-days for the same company. The features are then extracted from the data. Table 3 shows a list of input feature sets

Another important point we want to mention: because we do the daily prediction, the comments we use in the news published in day *T* should be before 9:00 am of day *(T+1)*. For example, if there is news about Google published on May 1st, 2009, then there may be some comments on that news published by the users after 9:00 am, May 2nd, 2009. We cannot use such comments since we do 1-day-ahead prediction.

TABLE 3 A list of input feature sets

No.	Feature Set Description
1	Stock Price, from *(T-6)* to *T*
2	Stock Volume, from *(T-6)* to *T*
3	Frequency of News, from *(T-6)* to *T*
4	Frequency of Comments, from *(T-6)* to *T*
5	Average Length/ Standard Deviation of Length of Comments, from *(T-6)* to *T*
6	Average /Standard Deviation of Response time of Comments, from *(T-6)* to *T*
7	Number of Comments Ranks, from *(T-6)* to *T*
8	Number of Loyals and Undefined, from *(T-6)* to *T*
9	Number of Early/Late Responder, from *(T-6)* to *T*

Our prediction is one day ahead, so that the output is the change rate of stock prices from day T to day $T+1$. The change rate is defined as:

$$C_r(T) = \frac{P(T+1) - P(T)}{P(T)} \qquad (5)$$

$C_r(T)$ is the change rate from T to $T+1$ and $P(T+1)$ is the stock closing price at day $T+1$.

C. Window-Shift Method for Training &Testing

We downloaded the web data from Engadget for the time from January 1, 2006 to August 17, 2008 for our three stocks. The reason why we choose this time span is that the news and comments before January 1, 2006 are not that many, and we cannot obtain some precise descriptions, e.g., the particular publishing time of comments after August 17, 2008.

In the experiments, we use a Window-Shift method for training and testing in our experiment since underlying dynamics may change rapidly in stock markets.

.

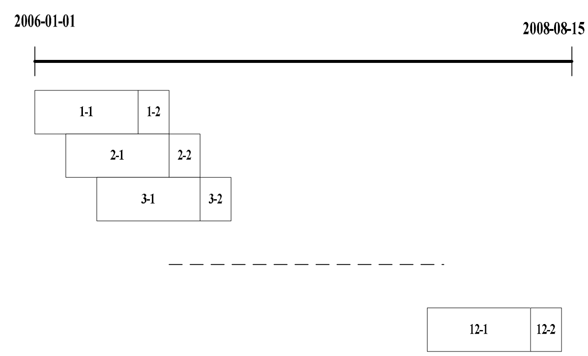

Figure 2: Window-Shift Method for Training &Testing

One training and corresponding testing period last for 8 months (nearly 160 trading days) and 2 months (nearly 40 trading days) respectively.

The N-1 (N=1, 2….12) in Fig. 2 means the N-th training period and the N-2 means the N-th testing period. There are totally 12 training and testing periods for each stock from January 1, 2006 to August 15, 2008.

V. Baseline Methods and Evaluation Methods

A. Baseline Methods

A list of proposed and baseline methods is shown in Table 4. MKL-All is our proposed method and others are baseline methods.

TABLE 4 A list of methods for experiments

Abbreviation	Description
MKL-All	Time Series + News + Comments , proposed MKL regression model
SVR-All	Time Series + News + Comments, SVR regression model
SVR-ts	Time Series Data Only, SVR regression model
SVR-news	News Frequency Only, SVR regression model
SVR-comments	Comments Dynamics Only, SVR regression model

B. Evaluation Methods

In the Performance Evaluation Component of our proposed model, we use three measures: MAE, MAPE and RMSE to evaluate the goodness of our proposed methods and baseline methods on magnitude prediction.

- **Mean Absolute Error (MAE)**

MAE is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error is given by

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P_i - A_i| = \frac{1}{n}\sum_{i=1}^{n}|E_i| \qquad (6)$$

where P_i is the forecast value, A_i is the actual value, and E_i is the absolute error.

- **Mean Absolute Percentage Error (MAPE)**

MAPE measures accuracy of fitted value relative to its actual values. It usually expresses accuracy as a percentage, and is defined by the formula:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{P_i - A_i}{A_i}\right| = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{E_i}{A_i}\right| \qquad (7)$$

- **Root Mean Square Error (RMSE)**

RMSE is a frequently-used measure of the differences between the values predicted by a model or an estimator and the values actually observed from the thing being modeled or estimated. It is defined by the formula:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}} \qquad (8)$$

VI. Experimental Results

A. MAE, MAPE, and RMSEs results

Fig. 3 shows an example of RMSEs results for the three target stocks. The X axis values (1 to 12) indicate the index of the testing period numbered as our Window-Shift method used for training and testing. We had 12 pairs of training and testing from January 1, 2006 to August 15, 2008 (The same holds for Fig. 4 in this paper.) The Y axis represents the value of RMSEs results for our proposed method and baseline methods.

Tables 5 to 7 show the average of the MAE, MAPE, and RMSE values in the 12 times testing periods for the three stocks. They indicate that our proposed model consistently (we just show one example in Fig. 3) and greatly outperforms other methods in terms of magnitude prediction which are evaluated by MAE, MAPE and RMSE.

TABLE 5 Average MAE, MAPE, RMSE of predictions for Amazon

	SVR-A	SVR-ts	SVR-n	SVR-c	MKL-A
MAE	1.5662	1.4448	1.5018	1.6072	1.0528
MAPE	2.4104	2.3085	2.3709	2.4623	1.5634
RMSE	1.9812	1.9214	1.9503	2.0199	1.6301

TABLE 6 Average MAE, MAPE, RMSE of predictions for Google

	SVR-A	SVR-ts	SVR-n	SVR-c	MKL-A
MAE	6.3900	7.3739	5.8763	6.7881	5.4716
MAPE	1.2138	1.3686	1.1180	1.2935	1.0468
RMSE	9.3357	10.505	9.0472	9.7988	8.5495

TABLE 7 Average MAE, MAPE, RMSE of predictions for Microsoft

	SVR-A	SVR-ts	SVR-n	SVR-c	MKL-A
MAE	1.0124	1.0078	0.9986	1.0222	0.2439
MAPE	3.3223	3.3276	3.2759	3.3430	0.8220
RMSE	1.0835	1.0788	1.0635	1.0984	0.3669

B. MKL coefficients

In the step of MKL training period, we obtained the coefficients for each feature set. Then we sum the coefficients for time series (stock price and volume), news (news frequency), and comments (from feature set 2 to feature set 7 in Fig. 1).

Fig. 4 shows the sum of coefficients for three company stocks in 12 times MKL training periods. For Microsoft, it shows that the coefficient for each source is stable within 12 MKL training periods and the coefficient for comments occupies nearly 70% of the total, which indicates that features for comments may contribute most to stock price movement prediction. For Amazon, the coefficients of news and comments change gradually but still the coefficients of comments occupy much more than other sources from 7th (starts in January 2007) to 11th (starts in September 2007 and ends in May 2008) training periods which roughly corresponds to the period when the stock price went up and kept high from April 2007 to January 2008 [22]. For Google, the changes of coefficients for comments are abrupt from 5th to 6th and 6th to 9th training period which roughly correspond to sudden surge and decline of stock prices around September 2007 and February 2008 respectively [23].

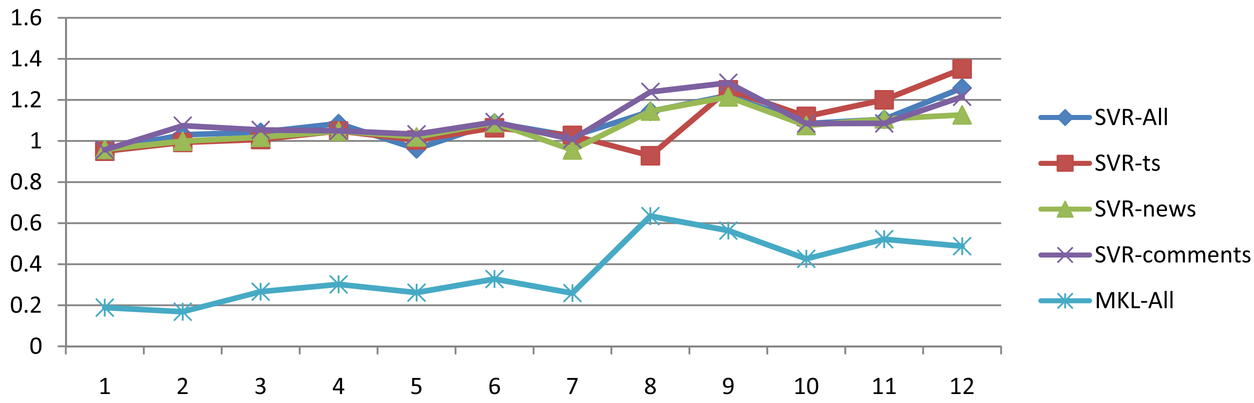

Figure 3. RMSEs results for Microsoft in the testing periods for proposed method and baseline methods

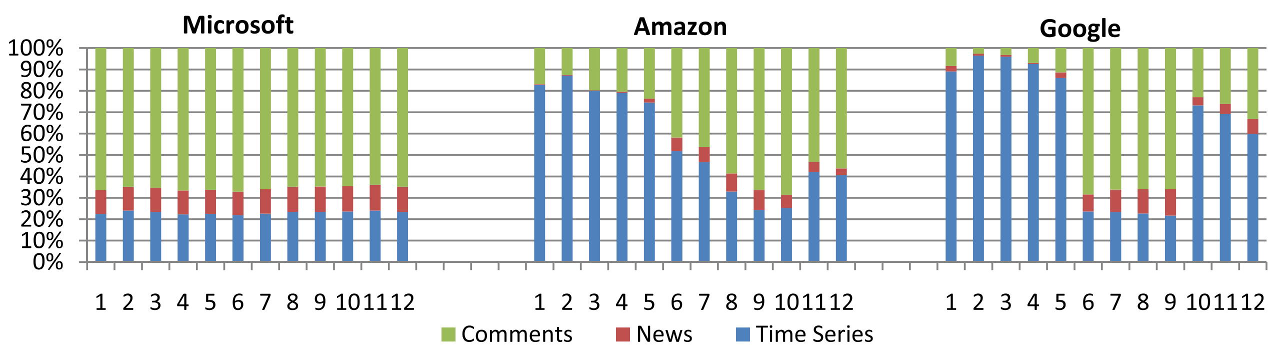

Figure 4. Coefficients results in the training step of Multiple Kernel Learning for three stocks

VII. Conclusion and Future Works

In this paper, we proposed a stock prediction model which includes Raw Data Preprocessing (RDP) component, Feature Selection (FS) component, MKL Stock Price Prediction (SP) component and Performance Evaluation (PE) component. First, the RDP component downloads data of time series, news, and comments; second, the FE component extracts features from the three different kinds of data sources; then, the SP component predicts the stock price based on a MKL regression framework; finally, the PE component evaluates the prediction results from MKL prediction component based on some magnitude evaluation measures. Results in Section VI show that our MKL prediction model outperforms other baseline methods.

In addition, coefficients of each source we obtained in the step of MKL training show us the possible relative correlation between stock prices and not only previous stock prices and volumes but also news frequency and numerical features of communication dynamics of comments on the news. Fig. 4 shows that in certain periods, the coefficients of news and comments for Amazon and Google increased (gradually or sharply); and in some other periods, the coefficients of news and comments decreased and time series data occupied most of the total coefficients. It may because that the abrupt surge or decline of stock prices for Google and Amazon [22, 23] during these periods indicates the popularity of Google and Amazon among the investors and users who post news or comments on Engadget. Moreover, we would imagine that when the price is predictable with good precision by time series data only, the coefficient for it is larger whereas when the price is predictable by time series data only with limited advances compared to SVRs the coefficients for it is smaller.

Since we only have research on news and comments of Engadget as the social networks source, changes on different social networks may give us different prediction results and MKL coefficients. Changing of the social network (e.g., Yahoo Finance Website) to investigate on the news and comments of different web sources is a future direction of our research.

Finally, the communication dynamics of news and comments in our research are mainly based on the works [11]. Extracting different kind of features of communication dynamics is another direction of future research. Moreover, in our research, we do not consider the text contents of news and comments. Combining text mining or sentiment analysis of news and comments with this research is another research direction.

ACKNOWLEDGMENT

This research was partially supported by Global-COE Program of Japan: “High-Level Global Cooperation for Leading Edge Platform on Access Spaces” of Keio University, Graduate School of Science and Technology.

REFERENCES

[1] J. Murphy, Technical analysis of the financial markets, Prentice Hall, London, 1998.
[2] C.J. Neely, Technical analysis and the profitability of U.S. foreign exchange intervention. Review, 1998, pp. 3–17.
[3] D. de la Fuente, A. Garrido, J. Laviada, A. Gomez, “Genetic Algorithms to Optimize the Time to Make Stock Market Investment.” GECCO 2006, Vol.2, pp.1857-1858, 2006
[4] F. Allen, R. Karjalainen, Using genetic algorithms to find technical trading rules, Journal of Financial Economics 51 (1999) 245–271.
[5] T.H. Hann, E. Steurer, Much ado about nothing? Exchange rate forecasting: neural networks vs. linear models using monthly and weekly data, Neurocomputing 10 (1996) 323–339.
[6] A.-S. Chen, M.T. Leung, Regression neural network for error correction in foreign exchange forecasting and trading, Computers and Operations Research 31 (2004) 1049–1068.
[7] J.T.-Y. Kwok, “The evidence framework applied to support vector machines,” IEEE Transactions on Neural Networks, vol.11, no.5, pp. 1162-1173, 2000.
[8] L.J. Cao, Support vector machines experts for time series forecasting, Neurocomputing 51 (2002) 321–339.
[9] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak, Andrew Tomkins, The predictive power of online chatter, Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, August 21-24, 2005 ,Chicago, Illinois, USA
[10] Munmun De Choudhury, Hari Sundaram, Ajita John, Doree Duncan Seligmann, Contextual Prediction of Communication Flow in Social Networks, Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, p.57-65, November 02-05, 2007
[11] Beibei Li , Shuting Xu , Jun Zhang, Enhancing clustering blog documents by utilizing author/reader comments, Proceedings of the 45th annual southeast regional conference, March 23-24, 2007, Winston-Salem, North Carolina
[12] M. D. Choudhury, H. Sundaram, A. John, and D. D. Seligmann, “Can blog communication dynamics be correlated with stock market activity?,” in ACM Hypertext 2008, Pittsburgh, PA, 2008, pp. 55-60.
[13] F.R.Bach, G.R.G. Lanckriet, M.I. Jordan, Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the 21st International Conference on Machine Learning (2004), pp. 6-13.
[14] T. Fletcher, Z. Hussain, J. Shawe-Taylor, Multiple kernel learning on the limit order book. JMLR: Workshop and Conference Proceedings 11 (2010) 167-174. Workshop on Applications of Pattern Analysis
[15] R. Luss, A. d’Aspremont, Predicting Abnormal Returns From News using Text Classification, arXiv:0809.2792v3, 2009
[16] Yeh, C.-Y., Huang, C.-W., Lee, S.-J., A multiple-kernel support vector regression approach for stock market price forecasting. Expert Systems with Applications (2010), doi:10.1016/j.eswa.2010.08.004
[17] Shangkun Deng, Kazuki Yoshyama, Takashi Mitsubuchi, Akito Sakurai. Hybrid Method of Multiple Kernel Learning and Genetic Algorithm for Forecasting Short Term Foreign Exchange Rate. Submitted to “Expert Systems With Applications”.
[18] V. Vapnik, The Nature of Statistical Learning Theory, Springer, NY, 1995
[19] S. Sonnenburg, G. Rätsch, C. Schäfer, B. Schölkopf, Large Scale Multiple Kernel Learning. Journal of Machine Learning Research (07 2006), 1531-1565
[20] Google Finance. http://finance.google.com/finance.
[21] Engadget. http://www.engadget.com/.
[22] Amazon stock price chart. http://www.google.com/finance?q=Amazon
[23] Google stock price chart. http://www.google.com/finance?q=google