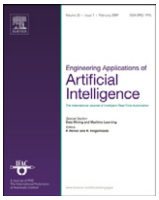
Engineering Applications of Artificial Intelligence 26 (2013) 1550–1561

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



# Stock price prediction based on a complex interrelation network of economic factors

Kanghee Park, Hyunjung Shin*

*Department of Industrial Engineering, Ajou University, San 5, Wonchun-dong, Yeoungtong-gu, 443-749 Suwon, South Korea*

ARTICLE INFO

*Article history:*
Received 28 February 2012
Received in revised form
24 September 2012
Accepted 30 January 2013
Available online 18 March 2013

*Keywords:*
Stock price prediction
Semi-supervised learning (SSL)
Machine learning

ABSTRACT

Stock price prediction is a field that has been continuously interesting. Stock prices are influenced by many factors such as oil prices, exchange rates, money interest rates, stock price indexes in other countries, and economic situations. Although these factors affect the stock price independently, they have an influence on the stock price through a complex interrelation, i.e., a network structure between these factors. In the stock prediction, the conventional methods represent limitations in reflecting the interrelation and complexity in these factors. In this paper, a stock prediction method using a semi-supervised learning (SSL) algorithm is proposed to circumvent such limitations. The SSL algorithm is a method that can implement a network consisting of nodes of the factors and edges of similarities between them. Through the network structure, the SSL algorithm is able to reflect the reciprocal and cyclic influences among the factors to prediction. The proposed model is applied to the stock price prediction from January 2007 to August 2008, using the global economic index and the stock prices of 200 individual companies listed to the KOSPI200.

## 1. Introduction

Interests on the stock price prediction have been continued according to the public understanding in stock investment. Many studies on stock price prediction employ various economic factors such as oil prices, exchange rates, interest rates, stock price indexes in other countries, domestic/global economic situations, etc. (Huang et al., 2008; Liu et al., 2009; Amilon, 2003; Chen et al., 1986). Methods for stock price prediction are diverse. Time series analysis is one of the most frequently adopted methods: Jeantheau (2004) predicted stock prices using an ARCH model, and Amilon (2003) and Liu et al. (2009) proposed a prediction method using a GARCH model based on the Skewed-GED Distribution for Chinese stock markets (Liu et al., 2009; Amilon, 2003). These methods assume that the future data will be varied as similar to those of the past. It is true, and a reasonable result can be obtained from time series analysis if given time series data are originated from natural phenomena, such as the numbers of sunspot, rain-falls, temperature, and so on. However, if data are given from economic or financial factors, it is difficult to expect reasonable prediction performance because of reciprocal and complex influences among the factors.

For example, Fig. 1 shows a case which considers KOSPI and Won-Dollar exchange rate to predict the stock price of Hyundai Motors. In Span A, the stock price of Hyundai Motors shows a similar pattern as rising and falling of exchange rate. In other words, this span is largely affected by exchange rate. However, in Spans B and D, the stock price of Hyundai Motors appears to be similar to a pattern of rising and falling in KOSPI rather than exchange rate. In the meantime, in Span C, it is difficult to find interrelations between KOSPI, exchange rate and the stock price of Hyundai Motors. This means that the stock price for Hyundai Motors is affected not only by these two indexes but also by other external indexes. Like this, stock price may be affected by various indexes. But influences by any specific index are not persistent. In addition, the influence by a stock price affects other financial–economic indexes and often returns to the stock price itself. This phenomenon also occurs for other financial–economic factors.

With time series analysis, however, there are methodological limitations to implement the reciprocal and complex influences among the factors into model formulation (Kim, 2003; Park et al., 2011). On the other hand, many studies on the stock price prediction have also been conducted in the machine learning domain. The artificial neural network (ANN) and support vector machine (SVM) methods have been frequently used as a typical model (Huang et al., 2005; Cao et al., 2005; Chen et al., 2003). Tay and Cao (2001) proposed a method that introduces financial time series data to the SVM, and Kanas (2003) attempted the prediction of the S&P500 index using the ANN model (Tay and Cao,

* Corresponding author.
*E-mail addresses:* can17@ajou.ac.kr (K. Park), shin@ajou.ac.kr (H. Shin).

0952-1976/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2013.01.009

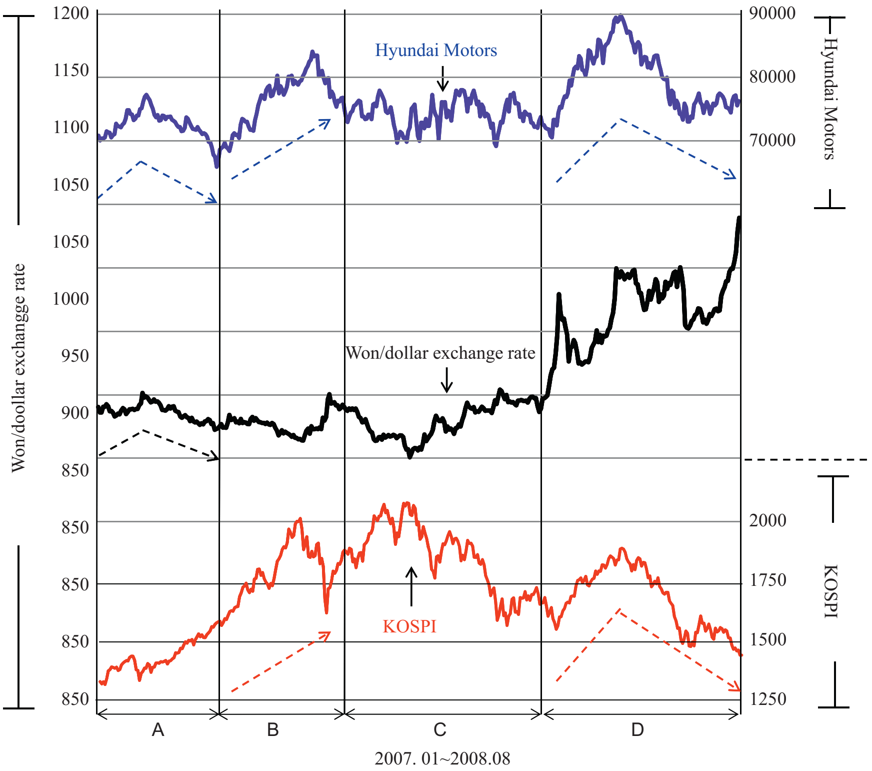


**Fig. 1.** A typical chart for reciprocal and complex influences among economic/financial factors (KOSPI Won/dollar exchange rates and stock prices of Hyundai motors, period: 2007. 01–2008. 08).

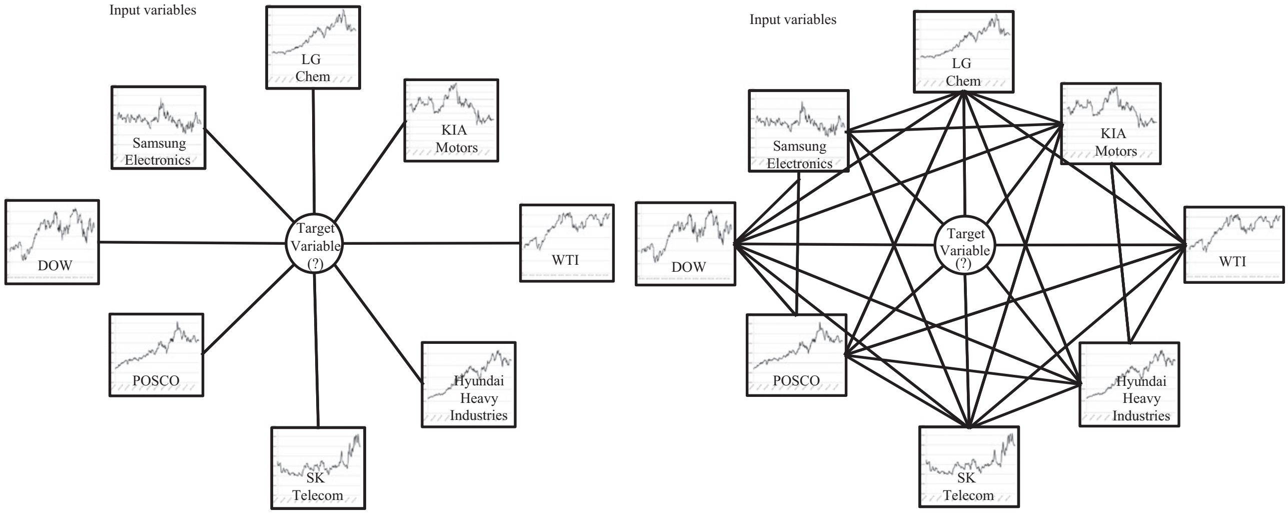


**Fig. 2.** Schematic description for the influence relations among factors by implementing approaches: (a) Time series analysis, SVM, and ANN methods, (b) SSL method.

2001; Kanas, 2003). Also, Yang et al. (2001) proposed an early warning system of commercial bank loan risks using the ANN model, and Bekiros and Georgoutsos (2008) analyzed that how uncertain news, which show a difficulty in identifying bullish and bearish factors, affect the NASDAQ index using the ANN model (Yang et al., 2001; Bekiros and Georgoutsos, 2008; Tsang et al., 2007). The methods using ANN and SVM may include the interrelation between the stock price and these factors in modeling. However it is still insufficient to explicitly formalize the mutual and complicated interrelation between the factors. As shown in Fig. 2(a), ANN and SVM can represent how interest rates, exchange rates, and oil prices affect stock prices primarily. However, it is somewhat difficult to express how the interests rates affect the exchange rates and then how these changes affect the next situation, i.e., how the second, third, and additional higher order interrelations affect the stock prices eventually. In addition, it is not easy to identify how the changes in stock prices caused by such a sequential process re-affect these factors later.

To solve this limitation, therefore, we propose to employ semi-supervised learning (SSL) which is a most recently emerged category of machine learning algorithms (Zhu, 2005; Shin et al., 2010). SSL defines the interrelation between factors to a network as illustrated in Fig. 2(b). SSL can consider the reciprocal and complex interrelation between factors based on a network. It connects individual factors via edges weighted by similarities between factors. Reciprocal and cyclic influences are delivered through the edges, and finally reach to the target factor—stock prices. The proposed SSL model is validated on the stock prices of 200 individual companies listed to KOSPI from January 2007 to August 2008.

This paper consists of five sections. Section 2 describes the methodology of SSL. Section 3 proposes a method of applying SSL to stock price prediction problem. Section 4 represents experiments and the results. Finally, Section 5 shows the conclusion of this study.

## 2. Semi-supervised learning (SSL)

Most recently, a category of machine learning algorithms, known as semi-supervised learning (SSL) has emerged, the main strength of which is that it allows taking advantage of the strengths of both supervised learning and unsupervised learning (Chapelle et al., 2006; Zhu, 2008). The primary goal of supervised learning is to build accurate classifiers or regressors using labeled data. On the other hand, unsupervised learning is usually employed to discover data structure from unlabeled data. In semi-supervised learning, meaningful representation of complicatedly structured data is identified from unlabeled data, and then the decision or regression function is achieved on both labeled and the unlabeled data, which is smooth with respect to the underlying geometry. SSL is regarded as a more pragmatic learning scheme since many practical domains are in such a situation that there is a large supply of unlabeled data but limited labeled data which can be expensive, difficult, and time-consuming to generate. Many related researches have shown validity of SSL in a number of application domains such as spam filtering (Zhu, 2005), document categorization (Shin and Tsuda, 2006), video surveillance (Shin et al., 2007), text classification (Subramanya and Bilmes, 2008), text chunking (Ando and Zhang, 2005), gene expression data classification (Gong and Chen, 2008; Bair and Tibshirani, 2004), webpage classification (Liu et al., 2006), etc. In those literatures, SSL is often compared with the representative models of supervised learning, and shows its superiority over them thanks to its capability of learning from only a few labeled data utilizing a large amount of unlabeled data. There has been a whole spectrum of interesting ideas on how to learn from both labeled and unlabeled data, e.g., the expectation-maximization based approach (Nigam et al., 1999), self-training (Yarowsky, 1995), co-training (Blum and Mitchell, 1998), Transductive support vector machines (Joachims, 1999), and the graph-based approaches such as graph mincuts (Blum and Chawla, 2001), harmonic approach (Zhu et al., 2003), local and global consistency (Zhou et al., 2004), etc. Among several types of SSL algorithms, a graph-based SSL is employed in our study (Shin et al., 2010; Shin et al., 2007).

In graph-based SSL algorithm, a data point $x_i \in R^M (i=1,\ldots, n)$ is represented as a node $i$ in a graph, and the relationship between data points is represented by an edge where the connection strength from each node $j$ to each other node $i$ is encoded as $w_{ij}$ of a weight matrix $W$ (Zhou et al., 2004). The labeled nodes have labels $y_l \in \{-1,1\} (l=1,\ldots,L)$, while the unlabeled nodes (question,?) have zeros $y_u=0$ $(u=L+1,\ldots,L+U)$. Fig. 3 presents a graphical representation of SSL.

A weight $w_{ij}$ can take a binary value (0 or 1) in the simplest case. Often, a Gaussian function of Euclidean distance between points with length scale $\sigma$ is used to specify connection strength:

$$w_{ij} = \begin{cases} \exp\left(-\frac{(x_i-x_j)^T(x_i-x_j)}{\sigma^2}\right) & \text{if } i\sim j \text{ ('}k\text{' nearest neighbors)} \\ 0 & \text{otherwise} \end{cases}, \tag{1}$$

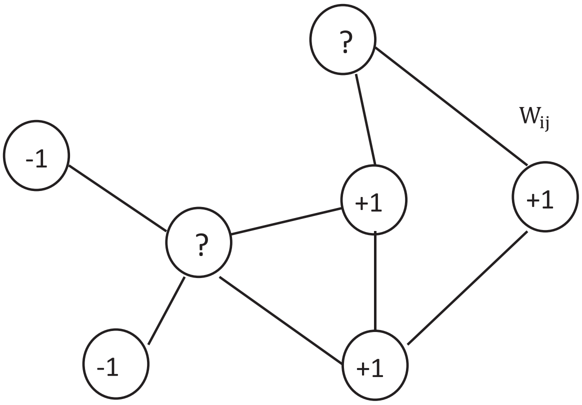


**Fig. 3.** Graph-based semi-supervised learning (SSL): The labeled node is denoted as "+1" or "−1", and the unlabeled node as "?".

Usually, an edge $i\sim j$ is established when node $i$ is one of $k$-nearest neighbors of node $j$ or node $i$ is within a certain Euclidean distance $r$, $\|x_i-x_j\|<r$. The algorithm will output an $n$-dimensional real-valued vector $f=[f_l^T\ f_u^T]^T=(f_1,\ldots,f_L,f_{L+1},\ldots,f_{L+U})^T$ which can be thresholded to make label predictions on $f_{L+1},\ldots,f_{L+U}$ after learning. It is assumed that $f_i$ should be close to the given label $y_i$ in labeled nodes (loss condition) and overall, $f_i$ should not be too different from its adjacent nodes $f_j$ (smoothness condition). One can obtain $f$ by minimizing the following quadratic functional:

$$\min_f \quad (f-y)^T(f-y)+\mu f^T L f, \tag{2}$$

where $y=(y_1,\ldots,y_l,0,\ldots,0)^T$, and the matrix **L**, called the graph Laplacian, is defined as $L=D-W$, $\mathrm{D}=\mathrm{diag}(\mathrm{d_i})$, and $d_i=\Sigma_j\omega_{ij}$. The first term corresponds to the loss function in terms of condition (a), and the second term represents the smoothness of the predicted outputs in terms of condition (b). The parameter $\mu$ represents trades between loss and smoothness. The solution to (2) is obtained as

$$f=(I+\mu L)^{-1}y, \tag{3}$$

where $I$ is the identity matrix.

## 3. Proposed method

To apply the graph-based SSL to time series prediction, we design a method of graph representation for time series data, and a procedure for obtaining predicted values from the graph. Assume that many stock prices are given as the input for the prediction problem for that of Hyundai Motors: the stock prices of LG Chem and KIA Motors, WTI intermediate oil price, other external factors, *etc*. To apply SSL to this problem, the original SSL graph in Fig. 3 is re-designed to a graph as in Fig. 4.

The nodes in the graph represent the factors that influence the stock price of Hyundai Motors, which are all time series variables. Then the edge between any two nodes $i$–$j$ stands for the similarity of the two sets of time series, represented as '$w_{ij} \in W$'. The label '$y_t$' on each node presents either 'up' (+1) or 'down' (−1) of the time series at time point $t$. In the graph of Fig. 4, the labels of Hyundai Motors are not known yet at time point $t$, and hence are unlabeled. To estimate the label $\mathrm{f_t}$, the similarity matrix of SSL was calculated at time point $t-1$, $W_{t-1}$.

At time $t$ information for the factors would not be enough to calculate similarities among them. Therefore, in order to make a prediction at the current time point, the similarity matrix is structured by using the dataset at the previous time point which holds the nearest and newest information. Fig. 5 represents the

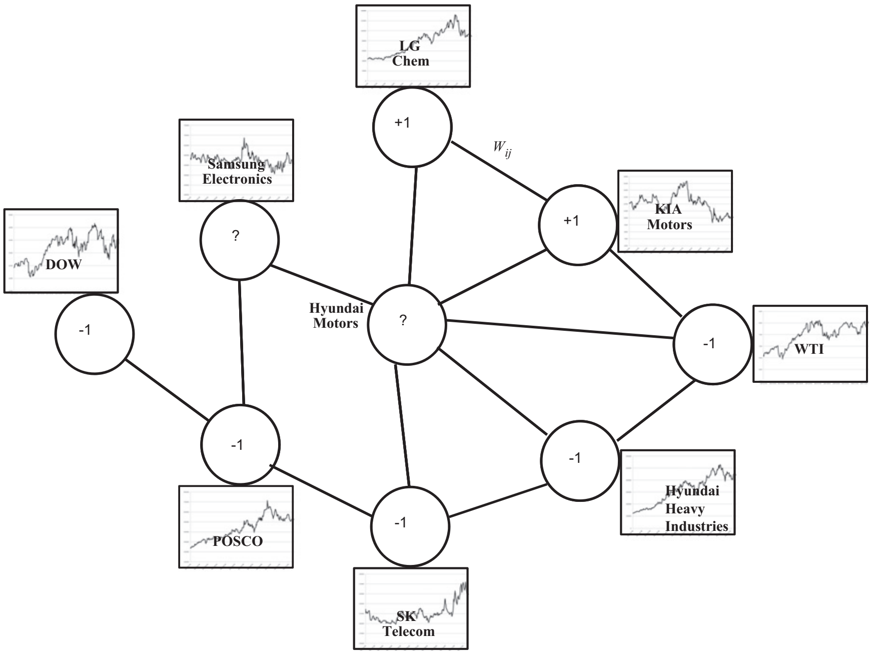


**Fig. 4.** Graph SSL representation for time series prediction.

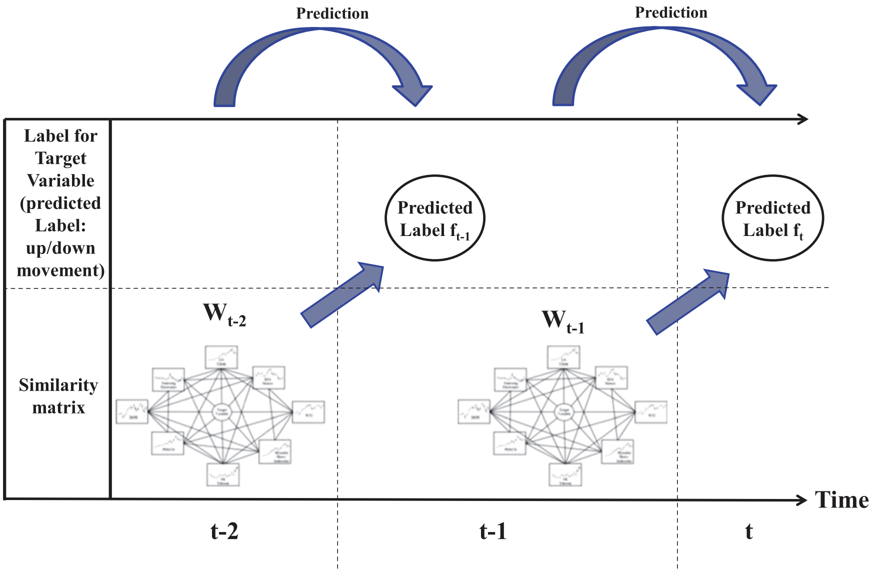


**Fig. 5.** Predicted value $f_t$ using $W_{t-1}$.

process of generating $W_{t-1}$, and using it for the prediction for the time point $t$.

Based on this set-up, we explain how to measure the similarity '$w_{ij}$' of a weight matrix $W$ and how to set the value for label '$y$'.

### 3.1. Similarity matrix

The design of the similarity matrix $W$ plays a critical part in the aspect of performance when using SSL (Shin et al., 2010, 2007). In the matrix $W$, each element represents how strongly the two nodes are related, with larger elemental value being associated with greater nodal similarity. In the proposed method, the time-series data are transformed by building technical indicators (TIs).

TIs are frequently used in financial forecasting as they offer the advantages of removing the noise (oscillatory noise) inherent in time series and illustrating the underlying structure, i.e., the tendencies and structural factors affecting variation (Kim, 2003, 2006; Park et al., 2011; Shin et al., 2007). Stock prices and other economic indices exist as time series data by the nature of the variables, and each of them is defined as a sequence as

$$X_t = \{x_1, x_2, \ldots, x_i, \ldots, x_t\}, \tag{4}$$

where $t$ represents the current time point, and $x_t$ is the corresponding value. The existence of $X_t$ as time series data induces several problems in the direct application of SSL to the data. As shown in Fig. 2, each of the nodes on the graph has its own time series, as shown in (4). For instance, the Hyundai Motors node has $X_t^{\text{Hyundai Motors}}$ and the LG chem mode also has $X_t^{\text{LGchem}}$. The problem is that it is difficult to draw the similarity between them directly from the two sets of series data. Therefore, individual time series are transformed into structural characteristics of time point $t$, i.e., $S_t^{\text{Hyundai Motors}}$and $S_t^{\text{LGchem}}$, representing the tendencies

**Table 1**
The definition of technical indicators (TIs).

| | TIs | Meaning |
|---|---|---|
| $s_1$ | $\mathrm{MA}_p(X_t)=\frac{1}{p}(x_t)+\frac{p-1}{p}\mathrm{MA}_p(X_{t-1})$ | $p$-moving average (exponential smoothing) |
| $s_2$ | $\mathrm{BIAS}_p(X_t)=\frac{x_t-\mathrm{MA}_p(X_t)}{\mathrm{MA}_p(X_t)}$ | The change rate of $x_t$ relative to $\mathrm{MA}_p(X_t)$ |
| $s_3$ | $\mathrm{OSC}_{p,q}(X_t)=\frac{\mathrm{MA}_p(X_t)-\mathrm{MA}_q(X_t)}{\mathrm{MA}_p(X_t)}$ | The change rate of $\mathrm{MA}_q(X_t)$ relative to $\mathrm{MA}_p(X_t)$ |
| $s_4$ | $\mathrm{ROC}_p(X_t)=\frac{x_t-x_{t-p}}{x_t}$ | The relative rate of change for $X_t$ between $p$ consecutive time points |
| $s_5$ | $K_t^p=\frac{x_t-\mathrm{Min}_{i=t-p-1}^{t}(x_i)}{\mathrm{Max}_{i=t-p-1}^{t}(x_i)-\mathrm{Min}_{i=t-p-t}^{t}(x_i)}$ | Standardization of $x_t$ |
| $s_6$ | $D_t^p=\mathrm{MA}_3(K_t^p)$ | 3-Moving Average of $K_t^p$ |
| $s_7$ | $\mathrm{RSI}_t^p=\frac{\sum_{i=t-p-1}^{t}(\lvert x_i-x_{i-1}\rvert)}{\sum_{i=t-p-t}^{t}(\lvert x_i-x_{i-1}\rvert)}$, | The relative strength index. |

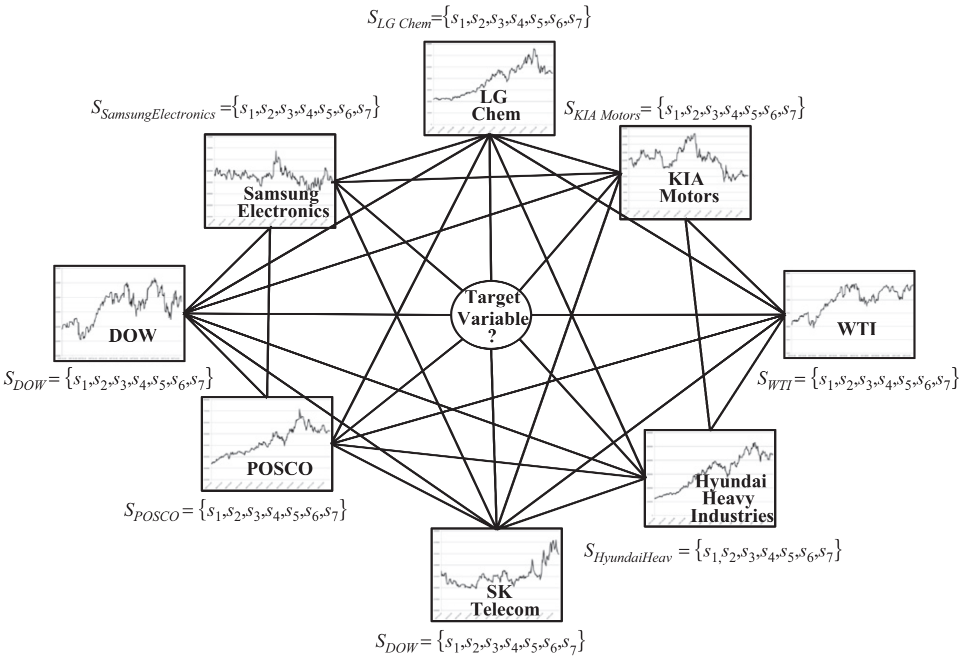

$S_{LG\ Chem}=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$
LG Chem
$S_{SamsungElectronics}=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$
Samsung Electronics
$S_{KIA\ Motors}=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$
KIA Motors
DOW
$S_{DOW}=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$
Target Variable ?
WTI
$S_{WTI}=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$
POSCO
$S_{POSCO}=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$
Hyundai Heavy Industries
$S_{HyundaiHeav}=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$
SK Telecom
$S_{DOW}=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$

**Fig. 6.** Similarity calculation using seven-tuple vectors.

and factors for variation of individual series. Table 1 summarizes the TIs used in this study. The similarity between the two nodes is measured by using the seven-tuple vector $S_t=\{s_1,s_2,s_3,s_4,s_5,s_6,s_7\}$ composed of MA, BIAS, OSC, ROC, K, D, and RSI. In other words, all nodes in Fig. 4 are represented by seven-tuple vectors, resulting in Fig. 6. The values of similarities between the nodes are calculated in the form of seven-tuple vector by following Eq. (1).

Using the TIs enables the time series data to be transformed into TIs-type data, while maintaining the time associations of the series, and thus eases their application to SSL.

### 3.2. Label

The label on the node in the SSL graph in Fig. 4 is designed to explain whether the predicted value of the corresponding variable is thumbs-up or down. It can be formulated as follows:

$$y_t=\mathrm{sign}(x_t-\mathrm{MA}_5(x_t)). \quad (5)$$

For instance, if the total amount of the Hyundai Motors' stock price ($t$) exceeds the 5-day moving average, (5) will give a '$y_t=+1$' label. On the contrary, the node is labeled as '$y_t=-1$' for the opposite case. And '$y_t=0$' if there is no information about the movement of the corresponding time series at time point $t$; the label is to be predicted. In the proposed method, we set the label of the target variable to '0'. Given label $y_t$, Eq. (3) provides the predicted value $f_t$ for every node, which can take on a real number unlike the values of label $y_t$.

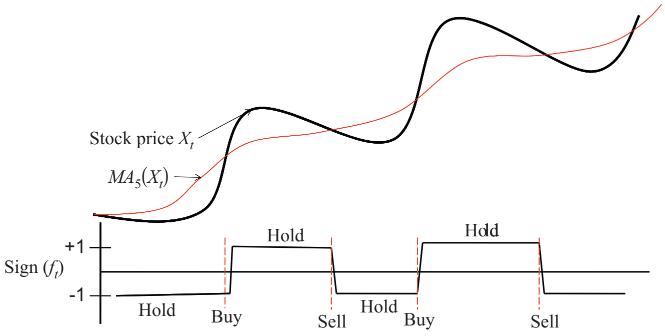


**Fig. 7.** Interpretation for forecasted values and simple trading strategy.

If '$f_t>0$', it means the stock price will increase relative to the average of the MA(5), therefore one can take the position of "buy" for the stock. On the other hand, one can take the position of “sell” otherwise. This procedure is described in Fig. 7.

### 3.3. Process summary

The proposed model is summarized as following. First, the stock price and financial–economical indexes are collected by the time *t*. Here, indexes whose values are known at time *t* are used for input variables and other unknown variables become target variables which are to be predicted. Second, each index is converted into seven-tuple vector by technical indicators transformation. Third, similarity matrix is structured by calculating Eq. (1) in the form of seven-tuple vectors. The matrix is calculated based on the dataset up to the previous time point (Section 3.1). Fourth, the labels of the indexes whose values are known at time *t* are calculated (Section 3.2). Finally, the up/down movements of the unknown indexes are predicted using the similarity matrix and the labels. Fig. 8 briefly shows the procedure mentioned above.

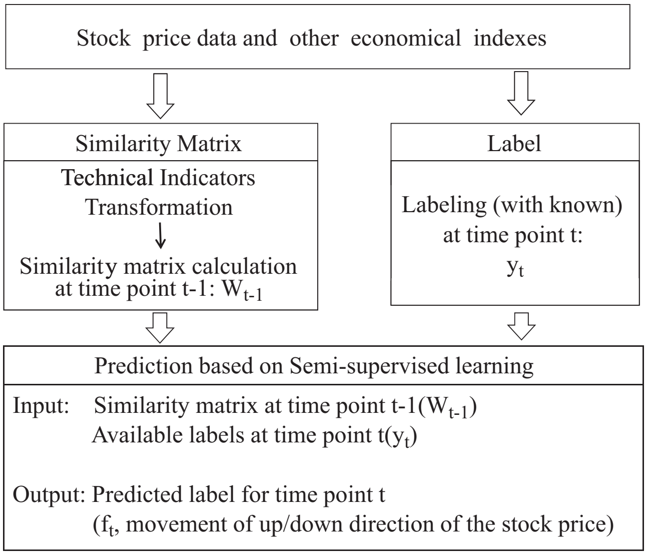

Stock price data and other economical indexes
Similarity Matrix
Technical Indicators Transformation
Similarity matrix calculation at time point t-1: $W_{t-1}$
Label
Labeling (with known) at time point t: $y_t$
Prediction based on Semi-supervised learning
Input: Similarity matrix at time point t-1($W_{t-1}$)
Available labels at time point t($y_t$)
Output: Predicted label for time point t
($f_t$, movement of up/down direction of the stock price)

**Fig. 8.** Procedure summary of stock price prediction using the SSL method.

## 4. Experiment

### 4.1. Data

The data used in this experiment was presented by a total of 403 daily data points from January 2007 to August 2008. The factors employed as variables were the major global economic indexes, such as Dow-Jones average (DOW), National association of securities dealers automated quotations (NASDAQ), Japanese stock market index (NIKKEI), Hang seng index (HSI), Shanghai composite index (SSE), Taiwan stock exchange corporation (TSEC), Financial times security exchange (FTSE), Deutscher aktien index (DAX), continuous assisted quotation index (CAC), Bombay stock exchange portmanteau of sensitive and index (BSE_SENSEX), Indice bovespa (IBOVESPA), Australia all ordinaries index (AORD), Korea composite stock price index (KOSPI), exchange rate (KRW-USD), the west Texas intermediate oil price (WTI), and the certificate of deposit (CD). Also, the stock prices of 200 companies listed to KOSPI200 were included. Table A1 (Appendix A) shows the list of these 200 companies. Fig. 9 shows a summary of these companies based on 18 different industry sectors.

A total number of 216 time series variables were employed in this study: 200 companies listed to KOSPI200 and 16 financial–economical indexes. Assuming that 16 financial–economical indexes were known at time t, they were used as input variables and the remaining individual stock prices were set as target variables. Note that similarities between the 216 indexes were calculated using the datasets up to the previous time point $t-1$.

### 4.2. Experimental setting

The SSL model proposed in this experiment was compared with the ANN and SVM models. The ANN model used a 3-layered MLP-structure. The SVM model used an RBF kernel function that has been known as an excellent performance model relatively. More details on SVM and ANN are described in Appendix B.

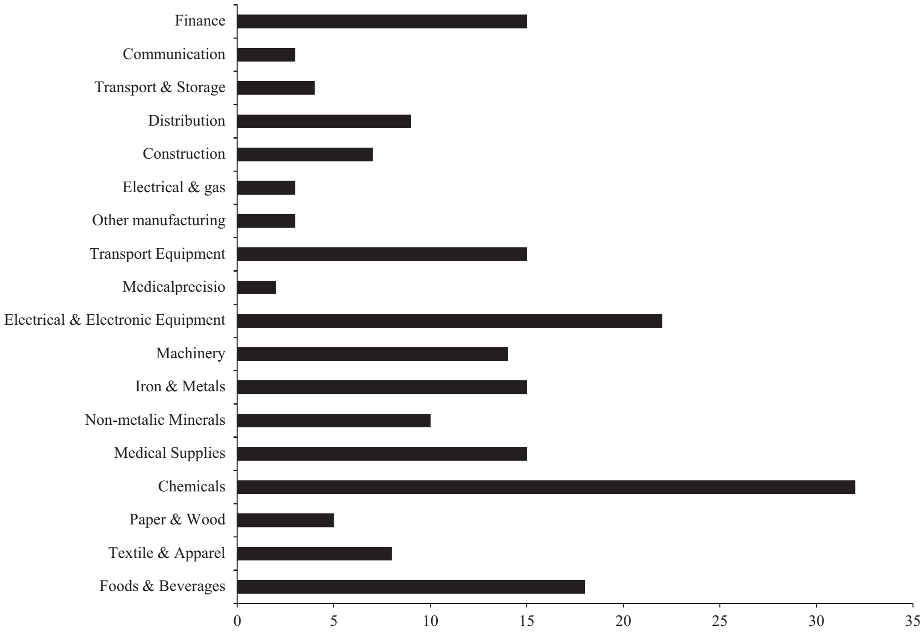


**Fig. 9.** Distribution of the 200 companies over industry sectors.

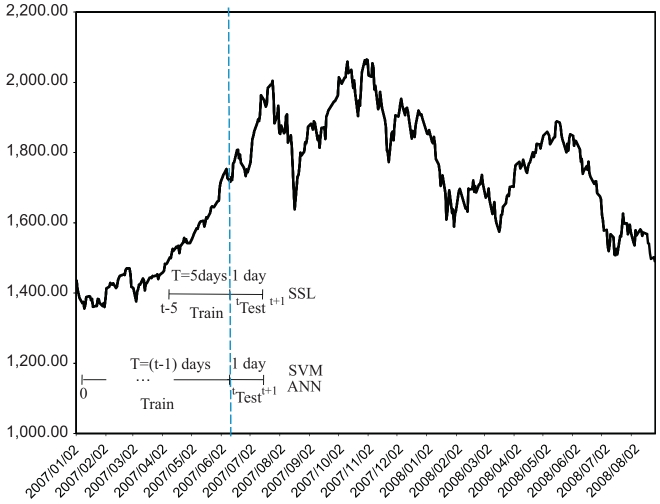


**Fig. 10.** Experimental setting.

A total of 403 daily data points were used for experiment. From January 2007 to May 2007, 103 daily data points were determined as training and validation sets. 20% of the random data points from the period were used for the specifications of the models and parameters. From June 2007 to August 2008, 300 daily data points were reserved for out-of-sample (test) evaluation and comparison of performances among the three models (Kim, 2003; Huang et al., 2005; Liu et al., 2009; Han and Kamber, 2006; Pai and Lin, 2005). For the test period, the performance of SSL was measured using a rolling forecast method which predicts a point of $t+1$ using the data from 1 to $t$, and similarly makes one-step-ahead predictions by moving forward the window (of size $t$) to the end of the test period (O'Connor et al., 2000). For the ANN and SVM models, the training set is very insufficient if the rolling forecast is applied. And both models are known to show better performance when the training set gets larger. Therefore, instead of removing the oldest data point from the training window, it was appended to the training set whenever the one-step-ahead prediction was made. As shown in Fig. 10, the training period employed in ANN and SVM was gradually increased, whereas that of SSL was fixed to the window size of $t$. The test period was dissected into 30 sections which have 10 test daily data points. The result will be shown for each of 30 test sections.

The parameters that are to be determined for the SSL model are k and μ and these parameters represent the number of neighbor node presented in Eq. (1) and the loss-smoothness tradeoff presented in Eq. (2), respectively. Also, the parameter values used in this experiment were determined as an optimal combination for the validation set presented in the range of $\{k,\mu\} \in \{2,3,4,5\} \times \{0.01,0.1,0.3,0.5,0.7,1,10,100\}$. In addition, the optimal value of the hidden node in ANN was determined in the range of {3–50} (Kim, 2003) and the parameters of kernel width (gamma) and misclassification tradeoff ($C$) in SVM were determined as an optimal combination of the values in the range of $\{\text{gamma}, C\} \in \{0.01,0.1,0.3,0.5,0.7,1,10,100\} \times \{0.01,0.1,0.3,0.5,0.7,1,10,100\}$ (Burges, 1998).

## 5. Results

### 5.1. Comparison of accuracy

To measure the prediction performance, the area under the curve (AUC), which is defined as the area under the receiver operating characteristic (ROC) curve, is used (Hanley and McNeil, 1982; Gribskov and Robinson, 1996). The ROC curve plots true positive rate as a function of false positive rate for differing classification thresholds as shown in Fig. 11. The AUC measures the overall quality of the model for all possible values of threshold rather than the quality at a single value of threshold. The closer the curve follows the left-hand border and then the top-border of the ROC space, the larger value of AUC the model produces; i.e., the more accurate the model is.

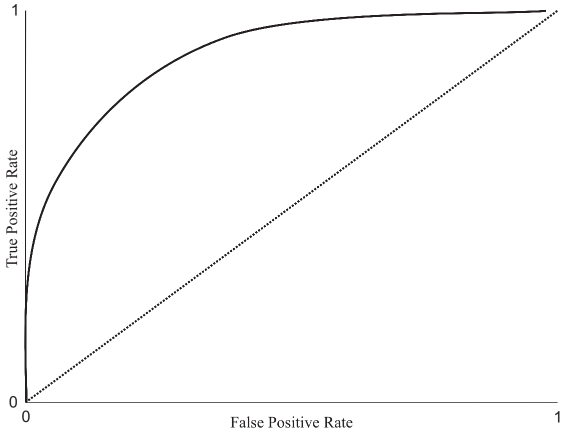


**Fig. 11.** ROC curve.

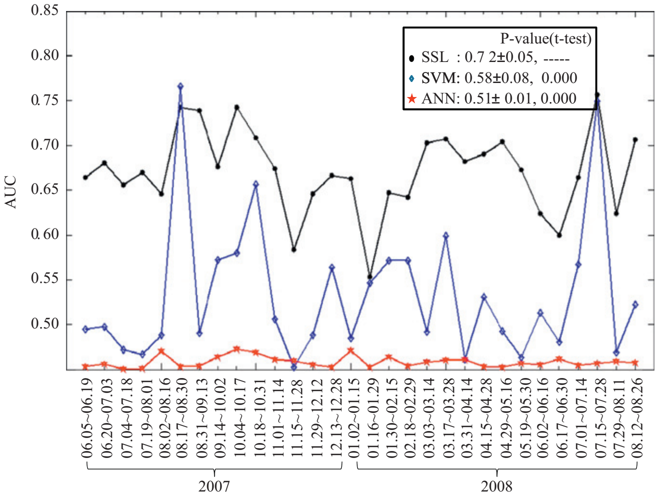


**Fig. 12.** AUCs comparisons with different methods.

Fig. 12 shows the graph of the values of AUC for the three models used in the 30 test sections. Points presented in the graph represent the average section values of AUC in which a section has 10 daily data points. The average AUC values in SVM and ANN for the total 30 sections were 0.58(±0.08) and 0.51(±0.01), respectively, but the value in SSL was 0.72(±0.05). Although ANN represented low volatility based on the standard deviation of 0.01, the AUC value was small compared to the other models. In the case of SVM, although it showed partially higher AUC values than SSL, it represented a very high deviation in its accuracy, whereas, SSL showed high accuracy in most sections compared to that of ANN and SVM. A $t$-test was applied to verify the significance that SSL represents better performance statistically than that of SVM and ANN. As a result, the difference in the performance between them showed statistical significance as shown in the upper right box in Fig. 12.

The outperformance of SSL model to other ones comes from its adaptable structure to changes in relations depending on time. In other words, the relations shown in Fig. 4 between financial–

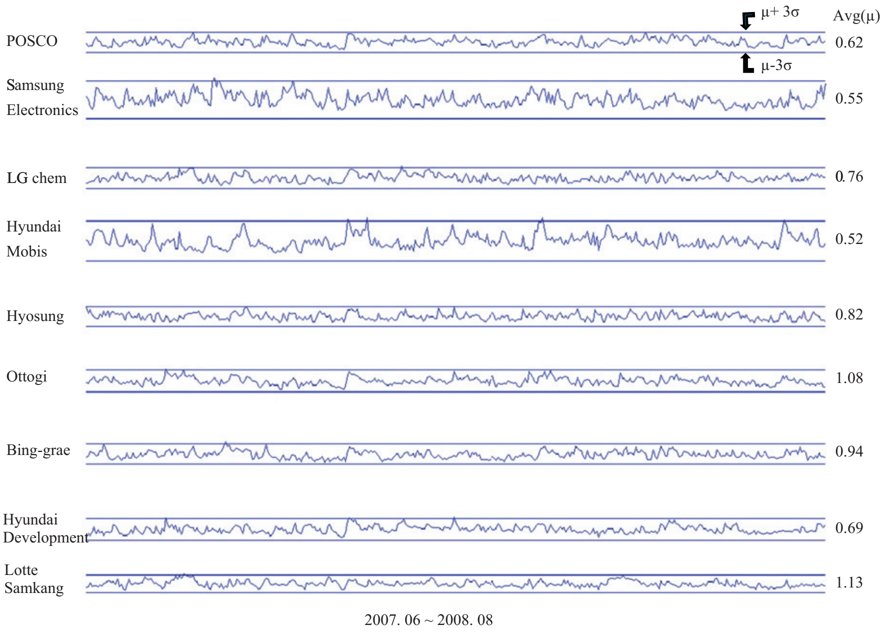


**Fig. 13.** Weight change between 30 test sections: Hyundai Motors vs. other companies.

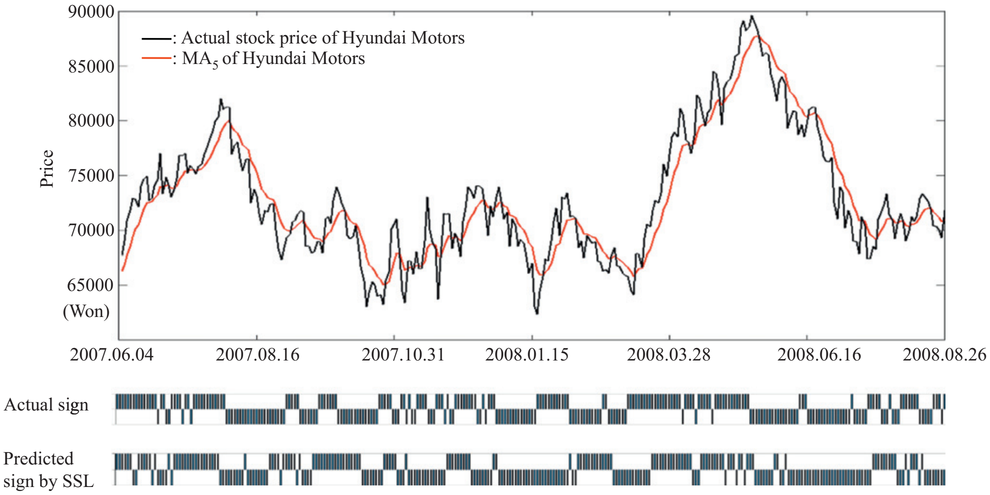


**Fig. 14.** *f* value and buy-and-sell strategy during the test period of June 2007 through August 2008.

economic factors may vary at each of the 300 time points and whenever it varies the similarity matrix for SSL reflects it. This is proved by looking into the immediate changes in weight values for the 300 time points. Fig. 13 exemplifies the typical changes in weight of 10 companies (out of 200 companies) including Hyundai Motors and other 9 companies, POSCO, Samsung Electronics, LG Chem, Hyundai Mobis, Hyosung, Ottogi, Bing-Grae Co., Hyundai Development, and Lotte Samkang Co. The value in the right side of the subfigure is a mean of weight ($\mu$) during the period while the upper and the lower lines indicate $\mu \pm 3\sigma$. As shown in the figure, the weight varies at every time point, but the width of the changes is bounded within $\mu \pm 3\sigma$. This implies that the relations are changing and yet stable. On the other hand, both ANN and SVM have difficulties to address the varying relations due to its innate model structure.

### 5.2. Comparison of profit

A stock price prediction model has to ensure both accuracy and earning rate while the stock price prediction model is on real investment (Barber et al., 2001). To calculate ROI, the benefit (return) of an investment is divided by the cost of the investment; the result is expressed as a percentage or a ratio. ROI measure is calculated by the following:

$$ROI = \frac{sell\ order\ price - buy\ order\ price}{buy\ order\ price} \times 100(\%).$$

Fig. 14 shows Hyundai Motors' buy and sell strategy as an example out of KOSPI200 listed companies. The upper panel in the figure shows the stock price and 5-day movement average value. Also, the lower actual sign represents the values calculated by Eq. (5) through reflecting the number of crosses in two lines. The lower panel shows actual sign and the predicted sign by SSL. The transaction was implemented according to such prediction values using the "one-point buy and sell strategy" presented in Fig. 7 (Jang and Lai, 1994). Table 2 shows the comparison of ROI according to the prediction values of the three models. ROI is calculated by each 30 test sections. Among the stock prices of 200 companies, the ROIs of the 10 companies mentioned above are presented (Table 2 shows only four companies' result and the

**Table 2**
The ROI values of four individual company stock prices for 30 test periods.

| Company | | Hyundai motors | | | | POSCO | | | | Samsung electronics | | | | LG chem | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Test period | SSL | SVM | ANN | BH | SSL | SVM | ANN | BH | SSL | SVM | ANN | BH | SSL | SVM | ANN | BH |
| 2007 | 06.05–06.19 | 6.53 | 0.00 | 0.00 | 0.00 | 4.28 | 5.88 | 1.21 | 0.00 | 2.88 | 6.83 | 3.96 | 0.00 | 2.21 | 0.00 | 0.00 | 0.00 |
| | 06.20–07.03 | 0.45 | 9.89 | 4.28 | 0.00 | 0.00 | 3.83 | 0.00 | 0.00 | 0.00 | −1.08 | 0.54 | 0.00 | 2.83 | 0.00 | 0.00 | 0.00 |
| | 07.04–07.18 | 0.00 | −0.89 | 2.07 | 0.00 | 0.00 | −2.21 | 0.00 | 0.00 | 0.00 | −0.72 | 0.00 | 0.00 | 3.33 | 6.38 | 0.00 | 0.00 |
| | 07.19–08.01 | 9.35 | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | −2.21 | 0.00 | 0.00 | −10.79 | 0.00 | 0.00 | 1.22 | 0.00 | 10.22 | 0.00 |
| | 08.02–08.16 | 1.31 | 2.51 | −6.65 | 0.00 | 8.56 | 0.00 | 5.81 | 0.00 | 10.97 | −5.22 | 6.83 | 0.00 | 0.00 | −1.12 | 0.00 | 0.00 |
| | 08.17–08.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 4.21 | −3.21 | 0.00 | −7.37 | −6.66 | −3.96 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 08.31–09.13 | −7.72 | −3.54 | 0.00 | 0.00 | −3.21 | 0.00 | −2.21 | 0.00 | 0.00 | 3.42 | −3.96 | 0.00 | 37.41 | 35.33 | 28.31 | 0.00 |
| | 09.14–10.02 | −0.15 | 0.00 | 0.00 | 0.00 | 0.00 | −5.83 | 4.22 | 0.00 | −3.78 | 6.30 | 2.70 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 10.04–10.17 | 0.00 | −2.51 | 0.00 | 0.00 | 0.00 | −2.35 | 0.00 | 0.00 | 0.00 | −5.04 | 0.00 | 0.00 | 0.00 | −2.22 | 0.00 | 0.00 |
| | 10.18–10.31 | −5.90 | 0.00 | 0.00 | 0.00 | −5.35 | 0.00 | −7.21 | 0.00 | −7.73 | 0.90 | −5.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 11.01–11.14 | −0.59 | −1.48 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 8.27 | 0.00 | 0.00 | 29.14 | 13.22 | 23.86 | 0.00 |
| | 11.15–11.28 | 0.00 | −13.74 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 9.71 | 0.00 | 0.00 | −2.21 | 2.21 | 0.00 | 0.00 |
| | 11.29–12.12 | 0.00 | −5.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | −6.12 | 0.00 | 0.00 | −1.22 | −3.21 | 0.00 | 0.00 |
| | 12.13–12.28 | 4.01 | −4.43 | 0.74 | 0.00 | 1.41 | 6.83 | 2.83 | 0.00 | 2.88 | −5.22 | 5.94 | 0.00 | −3.32 | −5.31 | 0.00 | 0.00 |
| 2008 | 01.02–01.15 | −8.01 | 0.00 | −11.82 | 0.00 | 0.00 | 0.00 | 1.41 | 0.00 | 0.00 | 2.52 | 0.00 | 0.00 | −18.31 | −19.33 | −26.22 | 0.00 |
| | 01.16–01.29 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.83 | 0.00 | 0.00 | 8.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 01.30–02.15 | −4.60 | 2.66 | 3.69 | 0.00 | 10.81 | 2.83 | 8.22 | 0.00 | 13.67 | 0.00 | 8.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 02.18–02.29 | −1.78 | 0.00 | −1.63 | 0.00 | −1.41 | 0.84 | −3.83 | 0.00 | −1.62 | −0.72 | 5.22 | 0.00 | −1.11 | 0.00 | 0.00 | 0.00 |
| | 03.03–03.14 | −0.15 | −2.51 | 0.30 | 0.00 | −2.82 | −4.83 | −2.21 | 0.00 | −2.70 | 10.61 | −5.94 | 0.00 | 0.81 | 3.34 | −0.21 | 0.00 |
| | 03.17–03.28 | 0.00 | 0.00 | 0.00 | 0.00 | 3.22 | 5.81 | 2.21 | 0.00 | 8.63 | 16.19 | 11.51 | 0.00 | 0.00 | 1.21 | 0.00 | 0.00 |
| | 03.31–04.14 | 0.00 | 15.95 | 15.66 | 0.00 | 6.83 | −1.81 | 3.35 | 0.00 | 7.19 | −5.22 | 6.83 | 0.00 | 5.52 | 8.83 | 4.22 | 0.00 |
| | 04.15–04.28 | 11.13 | 0.59 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 04.29–05.16 | 0.00 | 0.00 | 4.87 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | −5.04 | 0.00 | 0.00 | 0.00 | 1.21 | 0.00 | 0.00 |
| | 05.19–05.30 | 8.31 | 3.54 | −7.98 | 0.00 | 8.22 | 2.83 | 6.83 | 0.00 | 12.23 | −8.99 | 12.05 | 0.00 | −2.21 | −3.22 | −4.22 | 0.00 |
| | 06.02–06.16 | 0.00 | 0.00 | 0.89 | 0.00 | −6.88 | −0.42 | −8.31 | 0.00 | −2.34 | −5.94 | 0.54 | 0.00 | 5.31 | 2.38 | 3.32 | 0.00 |
| | 06.17–06.30 | 0.00 | −4.43 | −5.61 | 0.00 | −5.81 | −7.31 | −6.83 | 0.00 | −5.94 | 0.00 | −12.77 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 07.01–07.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | −7.31 | −1.21 | −5.31 | 0.00 |
| | 07.15–07.28 | 0.00 | −4.43 | 1.92 | 0.00 | −3.21 | −4.43 | −4.82 | 0.00 | −17.45 | −11.51 | −13.49 | 0.00 | −8.22 | −1.21 | −2.11 | 0.00 |
| | 07.29–08.11 | −3.12 | −5.91 | −1.33 | 0.00 | −4.22 | 0.00 | −3.22 | 0.00 | −2.88 | −4.68 | −3.24 | 0.00 | 0.00 | 0.00 | 5.96 | 0.00 |
| | 08.12–08.26 | 2.08 | −0.59 | −5.02 | 6.06 | −0.81 | −1.56 | −2.08 | 0.96 | −0.54 | −1.26 | −8.63 | −7.19 | −2.67 | 0.00 | 0.00 | 36.83 |
| Avg(%) | | **10.83** | −15.81 | −5.02 | 6.06 | **9.82** | 2.21 | −3.22 | 0.96 | **10.79** | −5.94 | 6.11 | −7.19 | **41.20** | 37.27 | 37.82 | 36.83 |

*—Trading commission: 0.015%.

**Table 3**
Average of 10 companies' ROI in 30 test sections.

| Model | SSL | SVM | ANN | BH |
|---|---|---|---|---|
| Avg of Hyundai Motors AUC | 10.83% | −15.81% | −5.02% | 6.06% |
| Avg of POSCO AUC | 9.82% | 2.21% | −3.22% | 0.96% |
| Avg of Samsung Electronics AUC | 10.79% | −5.94% | 6.11% | −7.19% |
| Avg of LG Chem AUC | 41.20% | 37.27% | 37.82% | 36.83% |
| Avg of Hyundai Mobis AUC | 23.67% | 9.21% | −6.17% | 9.11% |
| Avg of Hyosung AUC | 48.47% | 21.79% | 46.28% | 40.65% |
| Avg of Ottogi AUC | 34.69% | 17.97% | 11.33% | 27.63% |
| Avg of Bing-Grae Co AUC | −5.39% | −14.60% | −11.69% | −13.12% |
| Avg of Hyundai Development AUC | −18.35% | −36.27% | −36.63% | −38.84% |
| Avg of Lotte Samkang AUC | 7.61% | −3.97% | 2.57% | −14.29% |
| Avg of 10 companies AUC | 16.33% | 1.18% | 4.14% | 4.78% |
| The number of times of being the best profit model | 158 | 73 | 69 | -- |

other results are attached in Appendix C). In the table, the ROI of the Buy-and-Hold (BH) is used as a reference earning rate of which position is maintained without any transactions after the purchase at the start point. The lowest on-line commission 0.015% was considered as the trading commission when calculating ROI. Table 3 summarizes 10 companies' average ROI.

The average of 10 companies' earning rates of the SSL, SVM, ANN and the BH were +16.33%, 1.18%, 4.14%, and 4.78%, respectively. Comparing ROI by sections and companies, SSL showed the best ROI 158 times, SVM showed the best ROI 73 times, and ANN showed the best ROI 69 times. Thus, it was verified that SSL showed excellent performance in its earning rates.

## 6. Conclusions

In this study, a stock prediction method using time series data to SSL was proposed. The proposed method has the advantage that does not predict stock prices by considering the time series characteristics of the stock price in businesses like the conventional models but makes possible to predict the stock price using a network based on the fluctuation in other companies' stock prices and the economic index that affect the change in stock prices. Regarding the technical issue in the proposed method, the method used SSL and that leads to improve its predictability by including not only the influences on input variables and target variables but also the interrelation between input variables. Based on the combination of these advantages, it was possible to obtain the values of AUC and ROI as 0.72 and 16.33% respectively. The earning rate of the proposed method was compared using a "one-point buy and sell strategy" that has been considered as the simplest investment method. In future, it is expected that the earning rate will be improved as more various investment strategies are considered based on these results. In addition, the method proposed in this study can apply for predicting the fluctuation in stock prices for various stock items. Therefore, it is possible to expect profits and stabilities in investments as the results obtained in this study are combined with a portfolio optimization method.

## Acknowledgments

The authors would like to gratefully acknowledge support from Post Brain Korea 21 and the research Grant from National Research Foundation of the Korean Government (2010-0007804/2012-0000994).

## Appendix A

see Appendix Table A1.

## Appendix B

### *Artificial neural network (ANN)*

An ANN is an analytical system inspired by the structure of biological neural networks and their way of encoding and solving problems. We employed a well-analyzed and frequently used ANN architecture known as multi-layer perceptron with back-propagation algorithm. The ANN comprises of three types of layers: the input layer, hidden layers, and the output layer. The nodes in the input layer supply input signals (activation patterns from outside the system) to the nodes in the hidden layer via weighted connections. The overall result of the model is represented by the nodes in the output layer which send output signals (a weighted sum of the signals from the hidden nodes) on the basis of a transfer function. In ANN, the accuracy of the model often depends on the structure, i.e. the number of hidden nodes, and the initial weights associated with the connections between the nodes. Generally, the number of hidden nodes is selected by a trial-and-error fashion and the initial weights are randomly chosen. ANN is used to determine a set of weights w that minimize the total sum of squared errors:

$$E(w) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Note that the sum of squared errors depends on w because the predicted class is a function of the weights assigned to the hidden and output nodes.

**Table A1**
200 listed stock in KOSPI.

| | |
|---|---|
| Foods and beverages | Samyang Corporation, Hite Brew, Doosan Corporation, CJ Corp, Daehan Flour Mills Co., Daesang Corporation, Orion Corporation, Lotte Samkang Co., Namyang Dairy Product Co., Samyang Genex, Nong Shim Co., Lotte Confectionery Co., Bing-grae Co., Lotte Chilsung Beverage Co., Ottogi, Crown Confectionary, Dongwon F&B |
| Textile and apparel | Kyungbang Co., FnC Kolon Corp, Nasan, Handsome, The Basic House, LG Fashion Corporation |
| Paper and wood | Hankuk paper Mfg, Hansol Paper Co., Seha, Moorim paper. |
| Chemicals | Woongjin Chemical, Hankook Tire Co., Hanwha Co., Cheil Industries Inc, Kokon, Nexen Tire Co., KCC, Tae Kwang Industrial Co., Samsung Fine Chemicals Co., Hyosung SK Chemicals Co. Capro Korea PetroChemical Aekyung Petrochemical Youl Chon Chemical Hanwha Chemical OCI S-Oil Honam Petrochemical Korea Kumho Pertrochemical SKC UNID KPX Chemical, KPX Chemical Namhae Chemical, LG Household and Health Care, LG Chem, KP Chemical Corporation, Huchems Fine Chemical Corporation, Kumho Tires, Amore Pacific, Foosung |
| Medical supplies | Yuhan Corporation, Ildong Pharmaceutical Co., Dong A Pharmaceutical Co., JW Pharmaceutical, Chong Kun Dang Pharmaceutical, Bukwang Pharm, Ilsung Pharmaceuticals Co., Yungjin Pharm, Dong Wha Pharm, Green Cross, Il-Yang Pharm, Hanmi Pharmaceutical, Kwang Dong Pharm, LG Life Sciences Ltd., Daewoong Pharm |
| Non-metalic minerals | Chosun Refractories Co., Dongyang Mechatronics, Hanglas, Asia Cement Co., Hanil Cement Co., Ssangyong Cement Industrial Co., Sung Shin Cement Co., Samkwang Glass Ind Co., Hyundai Cement Co., Hankuk Glass Industries |
| Iron and metals | Young Poong Co., Dong Kuk Steel Mill Co., Se Ah Be steel Co., Kisco, Kiswire, SeAh Steel, Union Steel, Hyundai-Steel Co., BNG Steel Co., Posco, Poongsan, Korea Zinc, Hyundai Hysco, Dongbu Steel, Daehan Steel |
| Machinery | Dongbu Hannong Co., KC Cottrell, Shinsung ENG, DKME, Hyundai Elevator, Hankuk carbon, Halla Climate Control Co., Doosan Heavy Industries and Construction, Doosaninfracore, Hanmi Semiconductor, STX Engine, Sewon Cellontech, S&TC |
| Electrical and electronic equipment | Hynix Semiconductor Inc, Kumho Electric, Taihan, Daeduck GDS Co., Hansol Lcd Inc, Samyoung Electronics Co., Samsung Electronics, LS Industrial System, Samsung SDI, Daeduck Electronics, Korea Technology Industry, Samsung Engineering, LS, Celrun, Dongwon Systems, Iljin Electric, Korea electric terminal co, Sindoh, LG Display, Hyundai Autonet, LG Electronics |
| Medicalprecisio | Samsung Techwin, K.C. Tech |
| Transport equipment | Hyundai Motors, KIA Motors, Hanjin Heavy Industries, S&T Dynamics, Ssangyong Motor, Hyundai Heavy Industries, Samsung Heavy Industries, Hyundai Mipo Dockyard, Myungsung, Hyundai Mobis, Dongyang Mechatronics, Daewoo Shipbuilding & Marine Engineering, S&T Daewoo STX Offshore & Shipbuilding |
| Other manufacturing | Fursys, KT&G |
| Electrical and gas | Kepco, Kogas |
| Construction | Daelim Industrial Co., Hyundai Engineering and Construction Co., Kumho Industrial Co., GS Engineering and Construction, Hyundai Development, Daewoo E&C |
| Distribution and service | Samsung C&T, LG International, SK Networks Co, Amorepacific, LG, SK, Shinsegae Co, STX, S1, Dae Kyo, Coway, Lotte Shopping, Samsung Engineering, Cheil Worldwide Inc, SBS, Kangwon Land, NCsoft, Daewoo International, Hyundai Department Store, GS Holdings |
| Transport and storage | Hanjin Shipping Co., Korean Air Lines, Hyundai Merchant Marine |
| Communication | SK Telecom, KT, KTF |
| Finance | Samsung Fire and Marine Insurance, Hyundai Securities, Korean Reinsurance, Daegu Bank, Busan Bank, Woori Investment and Securities, Daewoo Securities, Samsung Securities, Industrial Bank of Korea, Mirae Asset Securities, Woori Finance Group, Shinhan Financial Group, Korea Investment Holdings, Hana Financial Group, KB |

**Table C1**
The ROI values of six individual company stock price for 30 test periods.

| Company | | Hyundai Mobis | | | | Hyosung | | | | Ottogi | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | | SSL | SVM | ANN | BH | SSL | SVM | ANN | BH | SSL | SVM | ANN | BH |
| 2007 | 06.05–06.19 | 6.11 | 3.22 | 6.99 | 0.00 | −4.39 | −2.68 | −1.34 | 0.00 | 0.00 | 0.00 | −0.11 | 0.00 |
| | 06.20–07.03 | −0.22 | −1.81 | −4.02 | 0.00 | −10.88 | −4.11 | −3.53 | 0.00 | 1.11 | 0.87 | 0.63 | 0.00 |
| | 07.04–07.18 | 0.00 | 1.31 | 1.12 | 0.00 | 0.00 | 3.17 | 8.59 | 0.00 | −0.33 | 0.00 | 0.00 | 0.00 |
| | 07.19–08.01 | 11.02 | 6.78 | 11.51 | 0.00 | 28.63 | 11.02 | 3.05 | 0.00 | 13.32 | 8.67 | 6.73 | 0.00 |
| | 08.02–08.16 | −3.31 | −5.54 | −13.99 | 0.00 | −14.30 | −16.86 | −21.21 | 0.00 | −3.90 | −5.31 | 0.00 | 0.00 |
| | 08.17–08.30 | 0.00 | −1.31 | −9.82 | 0.00 | 14.31 | −10.11 | −0.55 | 0.00 | 0.00 | −1.71 | 0.00 | 0.00 |
| | 08.31–09.13 | 5.81 | 0.00 | −0.36 | 0.00 | −1.15 | −11.11 | 0.00 | 0.00 | 13.28 | 0.00 | 5.31 | 0.00 |
| | 09.14–10.02 | −3.15 | 0.00 | 2.93 | 0.00 | 13.93 | 16.69 | 7.44 | 0.00 | 4.69 | 15.31 | 3.11 | 0.00 |
| | 10.04–10.17 | −2.87 | −1.84 | 0.00 | 0.00 | 0.00 | 0.00 | 2.86 | 0.00 | −5.23 | −6.73 | −6.11 | 0.00 |
| | 10.18–10.31 | 1.82 | 2.83 | −8.90 | 0.00 | −0.95 | 0.00 | 1.91 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 11.01–11.14 | −0.59 | −1.17 | 4.62 | 0.00 | 7.25 | 5.81 | 13.98 | 0.00 | −3.22 | 0.00 | −4.11 | 0.00 |
| | 11.15–11.28 | −0.22 | 0.00 | −6.40 | 0.00 | 0.00 | 0.00 | 2.10 | 0.00 | 8.33 | 3.11 | 0.00 | 0.00 |
| | 11.29–12.12 | 0.43 | 0.00 | 3.91 | 0.00 | 0.00 | 0.00 | 2.10 | 0.00 | −3.89 | 0.00 | −1.71 | 0.00 |
| | 12.13–12.28 | 0.22 | 0.00 | −0.47 | 0.00 | −18.13 | −5.31 | 0.76 | 0.00 | 1.71 | 0.87 | 0.00 | 0.00 |
| 2008 | 01.02–01.15 | −1.71 | −2.21 | −3.33 | 0.00 | 0.00 | 0.00 | −2.48 | 0.00 | 0.00 | 0.00 | 0.32 | 0.00 |
| | 01.16–01.29 | 0.00 | 0.00 | 0.83 | 0.00 | 0.00 | 0.00 | −9.88 | 0.00 | −6.71 | −7.81 | −7.81 | 0.00 |
| | 01.30–02.15 | 0.00 | 0.00 | −6.64 | 0.00 | 1.87 | 3.31 | 10.12 | 0.00 | 1.71 | 0.83 | 0.62 | 0.00 |
| | 02.18–02.29 | 0.00 | 0.00 | 0.00 | 0.00 | −0.95 | −2.35 | −0.95 | 0.00 | −5.33 | −6.71 | −6.11 | 0.00 |
| | 03.03–03.14 | 0.11 | −0.33 | 3.20 | 0.00 | −0.38 | −2.31 | −0.76 | 0.00 | −2.17 | 0.00 | 0.00 | 0.00 |
| | 03.17–03.28 | 0.00 | 0.00 | 2.61 | 0.00 | 5.15 | 2.21 | 0.00 | 0.00 | 0.00 | −2.27 | −3.22 | 0.00 |
| | 03.31–04.14 | 12.81 | 0.00 | −3.08 | 0.00 | 0.19 | −3.31 | −4.77 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 04.15–04.28 | 3.13 | 11.21 | 5.34 | 0.00 | 0.00 | 2.22 | 5.34 | 0.00 | 25.39 | 21.89 | 23.11 | 0.00 |
| | 04.29–05.16 | 7.12 | −1.31 | −1.78 | 0.00 | 0.00 | 12.36 | 19.47 | 0.00 | −3.13 | −4.17 | 0.00 | 0.00 |
| | 05.19–05.30 | −8.31 | −6.78 | 1.78 | 0.00 | 25.54 | −2.25 | −3.24 | 0.00 | 0.00 | 0.00 | −1.31 | 0.00 |
| | 06.02–06.16 | −3.53 | −4.21 | 6.00 | 0.00 | −0.76 | 1.21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | 06.17–06.30 | 2.21 | 3.83 | −1.67 | 0.00 | 7.25 | 6.30 | 6.30 | 0.00 | −0.31 | −1.22 | −0.71 | 0.00 |
| | 07.01–07.14 | −7.13 | 2.23 | 0.00 | 0.00 | −16.68 | 0.00 | 0.00 | 0.00 | −1.87 | −1.11 | −1.73 | 0.00 |
| | 07.15–07.28 | 1.81 | 2.83 | −0.12 | 0.00 | 6.11 | 11.10 | 14.50 | 0.00 | 3.22 | 3.22 | 3.22 | 0.00 |
| | 07.29–08.11 | 0.00 | 1.11 | 3.20 | 0.00 | −1.72 | 0.00 | 0.00 | 0.00 | −0.87 | 0.00 | 0.00 | 0.00 |
| | 08.12–08.26 | 2.11 | 1.78 | 0.69 | 9.11 | −1.53 | 0.00 | −2.29 | 40.65 | 0.00 | 1.11 | 1.73 | 27.63 |
| Avg(%) | | **23.67** | 9.21 | −6.17 | 9.11 | **48.47** | 21.79 | 46.28 | 40.65 | **34.69** | 17.97 | 11.33 | 27.63 |
| Company | | Bing-Grae Co | | | | Hyundai Development | | | | Lotte Samkang | | | |
| Model | | SSL | SVM | ANN | BH | SSL | SVM | ANN | B.H | SSL | SVM | ANN | BH |
| 2007 | 06.05–06.19 | 2.38 | 2.38 | 2.38 | 0 | −4.81 | −3.11 | −5.83 | 0 | −2.67 | −2.67 | −3.11 | 0 |
| | 06.20–07.03 | −0.87 | −1.11 | −0.87 | 0 | 0 | 0 | 0 | 0 | −5.36 | −6.11 | −6.11 | 0 |
| | 07.04–07.18 | 1.73 | 1.11 | 0 | 0 | 2.21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 07.19–08.01 | 0 | 0 | 0 | 0 | 1.71 | 2.31 | 1.11 | 0 | 14.29 | 12.28 | 11.83 | 0 |
| | 08.02–08.16 | −2.48 | −2.78 | 0 | 0 | 1.11 | 0 | 0 | 0 | −6.81 | −9.81 | −5.81 | 0 |
| | 08.17–08.30 | 1.31 | 0 | 0 | 0 | 1.31 | 3.17 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 08.31–09.13 | 0 | 1.11 | 1.31 | 0 | 0 | 0 | −0.87 | 0 | 20.09 | 15.34 | 17.31 | 0 |
| | 09.14–10.02 | 3.17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9.82 | 6.83 | 11.13 | 0 |
| | 10.04–10.17 | 0 | 0 | 0 | 0 | 16.89 | 8.91 | 13.89 | 0 | −8.31 | −5.31 | −11.83 | 0 |
| | 10.18–10.31 | −7.15 | −8.11 | −8.11 | 0 | 1.71 | 3.11 | 1.71 | 0 | −6.38 | −4.33 | −6.38 | 0 |
| | 11.01–11.14 | 0 | 0 | 0 | 0 | −22.17 | −27.71 | −25.87 | 0 | 0 | −5.31 | −4.12 | 0 |
| | 11.15–11.28 | 5.35 | 3.12 | 4.15 | 0 | 0 | 0 | 0 | 0 | 5.31 | 4.12 | 5.31 | 0 |
| | 11.29–12.12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −0.87 | 0 | 0 | 0 |
| | 12.13–12.28 | 3.32 | 0 | 0 | 0 | 14.19 | 14.19 | 14.19 | 0 | −5.11 | 0 | 0 | 0 |
| 2008 | 01.02–01.15 | 0 | 0 | 0 | 0 | −13.11 | −13.11 | −13.11 | 0 | −1.11 | −4.83 | −3.21 | 0 |
| | 01.16–01.29 | 0 | −0.87 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −5.83 | −3.38 | 0 |
| | 01.30–02.15 | −1.53 | −2.27 | −3.13 | 0 | 1.71 | 0.87 | 0.87 | 0 | 1.71 | 2.22 | 3.11 | 0 |
| | 02.18–02.29 | −7.15 | −9.13 | −8.13 | 0 | −1.31 | −2.11 | −2.31 | 0 | −0.51 | −1.71 | 0 | 0 |
| | 03.03–03.14 | 0 | 0 | 0 | 0 | −6.11 | −7.89 | −9.61 | 0 | −5.61 | −5.61 | −6.38 | 0 |
| | 03.17–03.28 | 2.13 | 1.71 | 3.12 | 0 | 0 | 0 | 0 | 0 | −3.33 | 0 | 0 | 0 |
| | 03.31–04.14 | 0 | 0 | 0 | 0 | 3.17 | 4.17 | 5.13 | 0 | 0 | −6.11 | 0 | 0 |
| | 04.15∼04.28 | 0 | 0 | 0 | 0 | −5.31 | −7.87 | −8.13 | 0 | 3.21 | 1.17 | 0.11 | 0 |
| | 04.29–05.16 | 1.71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 05.19–05.30 | −2.83 | 0 | 0 | 0 | −5.11 | 0 | 0 | 0 | −5.83 | 0 | 0 | 0 |
| | 06.02–06.16 | −0.87 | −0.71 | −1.11 | 0 | −3.11 | −12.31 | −10.83 | 0 | 1.21 | −3.11 | 0 | 0 |
| | 06.17–06.30 | 0.21 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −3.32 | 0 | −4.31 | 0 |
| | 07.01–07.14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −2.21 | 0 | −2.21 | 0 |
| | 07.15–07.28 | −2.31 | 0 | 0 | 0 | −6.13 | 0 | 0 | 0 | 0.87 | 3.21 | 0.87 | 0 |
| | 07.29–08.11 | 0 | 1.11 | 0 | 0 | 0 | −7.31 | −8.11 | 0 | 0 | 1.11 | 0 | 0 |
| | 08.12–08.26 | 0 | 1.11 | 0.21 | 0 | 0 | 5.31 | 5.31 | 0 | 0.53 | 1.71 | 0.53 | 0 |
| Avg(%) | | **−5.39** | −14.60 | −11.69 | −13.12 | **−18.35** | −36.27 | −36.63 | −38.84 | **7.61** | −3.97 | 2.57 | −14.29 |

*Support vector machine (SVM)*

SVM involves finding an optimal decision boundary i.e., maximizing the margin by finding the largest achievable distance among the separating hyperplane and the data points on either side. The classification can be represented by considering a set of input–output pairs $D==\{(x1,\ y1),\ (x2,\ y2),\ \ldots,\ (x\ell,\ y\ell)\}$, where $i=1,\ldots,\ \ell$. Here $x\in X$ and $y\in Y$ where '$Y$' represents the set of class labels, e.g., for binary classification $Y==\{-1,\ +1\}$. In a typical binary classification, training data points from two different classes are separated by a hyperplane. The separating hyperplane can be linear or non-linear. For linear classification, SVM computes the linear decision function in the central gap of the two classes by correctly classifying all the training data points and placing the decision function as far from the given data points as possible, to lessen the possibility of false prediction for the unseen data points. If classes are not linearly separable because of noisy data (measurement errors, uncertainty in class membership, etc.), we can still use the linear classifier with an error tolerance. In such a case, the aim is to find a balance between margin maximization and misclassification minimization. SVM solves the following quadratic programming problem to produce the maximum margin between the two classes:

$$\begin{aligned}&\min\Theta\,(\overrightarrow{w},\xi)=\frac{1}{2}\|\overrightarrow{w}\|^2+C\sum_{i}^{M}\xi_i\,,\\&\quad s.t.\quad y_i(\overrightarrow{w}\,.\,\Phi(\overrightarrow{x}_i)+b)\geq 1-\xi_{i,}\\&\quad \xi_i\geq 0,i=1,\ldots\ldots,M.\end{aligned}\tag{B1}$$

The parameter $C$ in Eq. (B1) is the penalty for misclassifying a data point. The higher the value of $C$, the more the SVM training is compelled to avoid classification errors. The parameter $\xi_i$ is the non-negative slack variable, which allows a certain level of misclassification for an inseparable case. If the data points are separated by a non-linear hyperplane because of some intrinsic property of the problem, it is more appropriate to map the input feature space to a high-dimensional feature space where the data points are separated by a linear hyperplane. This mapping process $\varphi$ is conducted by kernel functions. Among many types of kernel functions, the RBF kernel $k(\underset{u}{\rightarrow},\underset{v}{\rightarrow})=e^{-\gamma|\underset{u}{\rightarrow}\,\underset{v}{\rightarrow}|^2}$ is most widely used. The parameter values of the tradeoff $C$ and the kernel width $\gamma$ are specified by users, and affect the performance of SVM.

## Appendix C

see Appendix Table C1.

## References

Amilon, H., 2003. GARCH estimation and discrete stock prices: an application to low-priced Australian stocks. Econ. Lett. 81, 215–222.

Ando, R.K., Zhang, T., 2005. A High-Performance Semi-Supervised Learning Method for Text Chunking. In: ACL '05 Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Ann Arbor, Michigan, pp. 1–9.

Bair, E., Tibshirani, R., 2004. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol. 2, 511–522.

Barber, B., Lehavy, R., Mcnichols, M., Trueman, B., 2001. Can Investors profit from the prophets? Security analyst recommendations and stock returns. J. Finance 1, 531–563.

Bekiros, S., Georgoutsos, D., 2008. Direction-of-change forecasting using a volatility-based recurrent neural network. J. Forecast. 27, 407–417.

Blum, A., Chawla, S., 2001. Learning from labeled and unlabeled data using graph mincuts. In: ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning San Francisco, pp. 19–26.

Blum, A. and Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In: COLT' 98 Proceedings of the Eleventh Annual Conference on Computational Learning Theory, New York, pp. 92–100.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discovery 2, 121–167.

Cao, Q., Leggio, K.B., Schniederjans, M.J., 2005. A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. Comput. Oper. Res. 32, 2499–2512.

Chapelle, O., Scholkopf, B., Zien, A., 2006. Semi-Supervised Learning. MIT Press, Cambridge, England.

Chen, A.-S., Leung, M.T., Daouk, H., 2003. Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index. Comput. Oper. Res. 30, 901–923.

Chen, N.-F., Roll, R., Ross, S.A., 1986. Economic forces and the stock market. J. Bus. 59, 383–403.

Gong, Y.-C., Chen, C.-L., 2008. Semi-supervised method for gene expression data classification with Gaussian fields and harmonic functions. In: 19th International Conference on Pattern Recognition(ICPR 2008), Tampa, FL, pp. 1–4.

Gribskov, M., Robinson, N.L., 1996. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Comput. Chem. 20, 25–33.

Han, J., Kamber, M., 2006. Data Mining—Concepts and Techniques, 2nd edition Morgan Kaufmann.

Hanley, J.A., McNeil, B., 1982. The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. Radiology 143, 29–36.

Huang, C.-J., Yang, D.-X., Chuang, Y.-T., 2008. Application of wrapper approach and composite classifier to the stock trend prediction. Expert Syst. Appl. 34, 2870–2878.

Huang, W., Nakamori, Y., Wang, S.-Y., 2005. Forecasting stock market movement direction with support vector machine. Comput. Oper. Res. 32, 2513–2522.

Jang, G.-S., Lai, F., 1994. Intelligent trading of an emerging market. In: Deboeck, G.J. (Ed.), Trading on the Edge. John Wiley & Sons, Inc., pp. 80–101.

Joachims, T., 1999. Transductive inference for text classification using support vector machines. In: International Conference on Machine Learning, San Francisco, pp. 200–209.

Kanas, A., 2003. Non-linear forecasts of stock returns. J. Forecast. 22, 299–315.

Kim, K.-J., 2003. Financial time series forecasting using supportn vector machines. Neurocomputing 55, 307–319.

Kim, K.-J., 2006. Artificial neural networks with evolutionary instance selection for financial forecasting. Expert Syst. Appl. 30, 519–526.

Liu, H.-C., Lee, Y.-H., Lee, M.-C., 2009. Forecasting china stock markets volatility via GARCH models under skewed-GED distribution. Journal Money Investment Banking, 5–14.

Liu, Q., Sung, A.H., Chen, Z., Liu, J., Huang, X., Deng, Y., 2009. Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. MAQC-II Gene Expression 4, 1–24.

Liu, R., Zhou, J., Liu, M., 2006. A graph-based semi-supervised learning algorithm for web page classification. In: International Conference on Intelligent Systems Design and Applications, China, pp. 856–860.

Marcus, O'Connor., William, Remus., Kenneth, Griggs., 2000. Does updating judgmental forecasts improve forecast accuracy? Int. J. Forecast. 16, 101–109.

Nigam, K., Mccallum, A.K., Thrun, S., Mitchell, T., 1999. Text classication from labeled and unlabeled documents using EM. Mach. Learn. 39, 1–34.

Pai, P.-F., Lin, C.-S., 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33, 497–505.

Park, K., Hou, T., Shin, H., 2011. Oil price forecasting based on machine learning techniques. J. Korean Inst. Ind. Eng. 37, 64–73.

Shin, H., Tsuda, K., 2006. Prediction of protein function from networks. In: Chapelle, O., et al. (Eds.), Semi-Supervised Learning. MIT press, pp. 339–352.

Shin, H., Lisewski, A.M., Lichtarge, O., 2007. Graph sharpening plus graph integration: a synergy that improves protein functional classification. Bioinformatics 23, 3217–3224.

Shin, H., Hill, N.J., Lisewski, A.M., Park, J.-S., 2010. Graph sharpening. Expert Syst. Appl. 37, 7870–7879.

Subramanya, A., Bilmes, J., 2008. Soft-supervised learning for text classification. in EMNLP '08. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, pp. 1090–1099.

Tay, F.E.H., Cao, L., 2001. Application of support vector machines in financial time series forecasting. Omega 29, 309–317.

Thierry, Jeantheau, 2004. A link between complete models with stochastic volatility and ARCH models. Finance and Stochastics 8, 111–131.

Tsang, P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong, K., Wong, T.-L., 2007. Design and implementation of NN5 for Hong Kong stock price forecasting. Eng. Appl. Artif. Intell. 20, 453–461.

Yang, B., Li, L.X., Xu, J., 2001. An early warning system for loan risk assessment using artificial neural networks. Knowl.-Based Syst. 14, 303–306.

Yarowsky, D., 1995. Unsupervised word sense disambiguation rivaling supervised methods. In: ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Stroudsburg, pp. 189–196.

Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B., 2004. Learning with local and global consistency. In: Advances in Neural Information Processing Systems 16(NIPS), Whistler, British Columbia, pp. 321–328.

Zhu, X., 2005. Semi-Supervised Learning with Graphs. Carnegie Mellon, Pittsburgh, PA 15213.

Zhu, X., 2008. Semi-Supervised Learning Literature Survey. Technical Report TR-1530, Wisconsin-Madson.

Zhu, X., Ghahramani, Z., Lafferty, J., 2003. Semi-supervised learning using Gaussian fields and harmonic functions. In: International Conference on Machine Learning(ICML2003), Washington DC, pp. 912–919.