Pattern graph analysis-based stock price prediction on big stock data

Seungwoo Jeon[1], Bonghee Hong[2,*], Victor Chang[3]

Dept. of Electrical and Computer Engineering

Pusan National University, Busan, South Korea

Abstract

Stock price prediction is the most difficult field due to irregularity. However, because stock price is sometimes showing similar patterns and is determined by a variety of factors, our new idea is to find similar patterns in historical stock data to achieve daily stock price with high prediction accuracy and potential rules selecting main factors that have significant effect on the price among all factors simultaneously. The goal of our paper is to suggest a new complex methodology that finds the optimal historical dataset with similar patterns according to various algorithms for each stock item and provides a more accurate prediction of daily stock price. First, we use a Dynamic Time Warping algorithm to find patterns with the most closely similar situation adjacent to a current pattern. Second, we select the determinant that are most influenced by the stock price using feature selection based on Stepwise Regression Analysis. Moreover, we generate an artificial neural network model with selected features as training data for predicting the best stock price. Finally, we use Jaro-Winkler distance with Symbolic Aggregate approXimation (SAX) as prediction accuracy measure to verify our model.

Keywords: Stock price prediction, Dynamic time warping, Feature selection, Artificial neural network, Jaro-Winkler distance, Symbolic Aggregate approXimation

*Corresponding author

Email addresses: i2825t@pusan.ac.kr (Seungwoo Jeon), bhhong@pusan.ac.kr (Bonghee Hong), ic.victor.chang@gmail.com (Victor Chang)

[1]Post Doctoral Researcher

[2]Professor

[3]Professor

1. Introduction

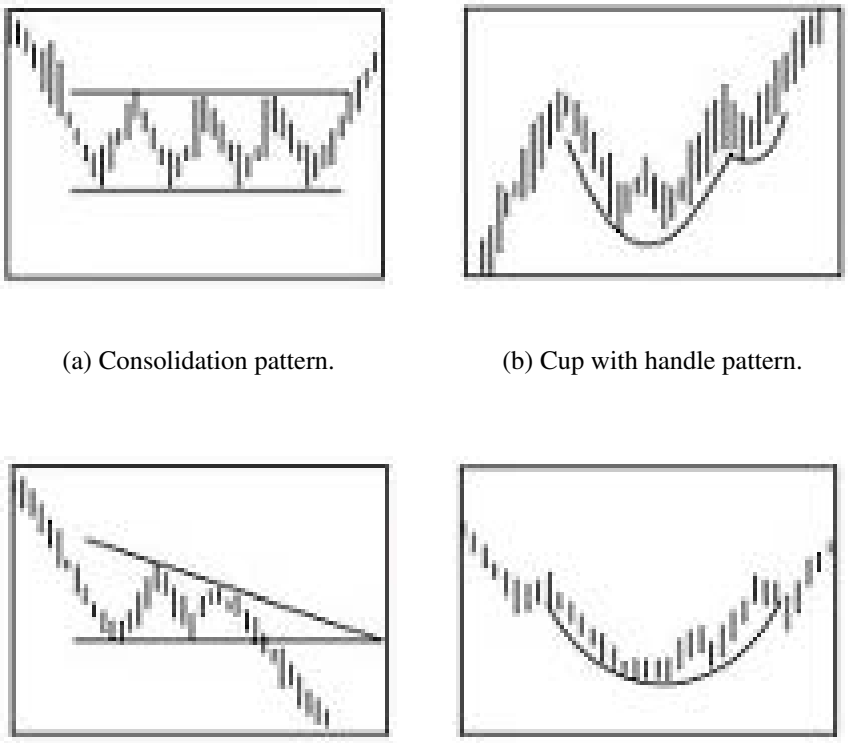
Stock price provided by the KOSCOM consists of thirty-two items (four groups: domestic and foreign or buying and selling) such as domestic selling high price, foreign selling opening price and domestic buying completion amount. For example, even if there are stock prices with the same value, its inside combination is different. Namely, domestic selling high price is downturn and domestic buying completion amount is upturn whereas domestic selling high price is upturn and domestic buying completion amount is downturn. Because of very changeable items, the goal is to predict the next stock price pattern graph using them and it would be of value.

Analysis and prediction in stock market are being studied using various methods such as machine learning and text mining. First of all, as data mining studies using daily stock data, there are prediction researches based on support vector machine (SVM) ([7, 22]) to determine whether the new pattern data belongs to a certain pattern category, artificial neural network (ANN) ([26, 27]) to have good prediction even if complex relationship of variables, and autoregressive integrated moving average (ARIMA) ([35, 40]) to identify and predict time series variation. Unlike machine learning, there are several prediction researches based on word analysis of news articles ([33, 34, 25]).

Since these researches have predicted daily stock prices using daily closing price, it is not enough make predictions in a short period of time such as one hour and 30 minutes. Moreover, even if they have analyzed significance of variables and increased the prediction accuracy of the model through eliminating the unimportant variables, error rates of the prediction are higher due to the use of any data contained in the outliers data.

The stock price consists of several patterns such as consolidation, cup with handle, double bottom, and saucer, as shown in Figure 1. Since these patterns appear repeatedly at time intervals, if we find a parallel pattern to the current pattern, it will be able to predict the following pattern.

Focusing on this point, we propose a new method for generating stock prices prediction based on historical stock big data in this paper. First, unlike existing studies that mostly use closing price data, we use tick by tick data for short term prediction and aggregate them to transform non-continuous data to continuous data. Then, we make some patterns similar to the current pattern through a dynamic time warping algorithm and select important features affecting the stock



(a) Consolidation pattern.

(b) Cup with handle pattern.

(c) Double bottom.

(d) Saucer.

Figure 1: Various stock patterns [6, 43].

price by using stepwise regression with them. Finally, we generate an artificial neural network using data to be completed similar patterns and feature selection as input data for high predictive accuracy through learning to derive the best results.

Because the pattern size of the stock price is not fixed, it may appear in a short time or for a long time next time and the upper and lower sizes may be smaller or larger. In other words, it is necessary to find out singular points of the day for easily comparing the predicted graph and the actual graph. Therefore, to find singular points of the day, we use a prediction accuracy measure combined Symbolic Aggregate approXimation (SAX) [31] and Jaro-Winkler distance [41]. This is to recognize the similarity(accuracy) between the predicted graph and the actual graph using the strings transformed from two approximated graphs

Thus, we propose prediction system based on big data processing (Hadoop, Hive, RHive) and analysis (R) tools for next stock price prediction. The system composed of four connected computers includes five steps. As preprocessing, first step is to transform tick by tick data to aggregated data at five minutes intervals to facilitate the prediction and make daily patterns with five minutes generation unit using Hadoop and RHive query. Second step is to find all similar patterns for three months using dynamic time warping algorithm provided by R function. Then, the system repeatedly remove insignificant variables through stepwise regression on R function. Next, the system uses artificial neural network to generate final prediction model according to numerous simulations.

The main contributions of this paper can be summarized as follows.

- We generate a prediction model for the stock prices by applying artificial neural network through dynamic time warping as pattern matching algorithm and stepwise regression as significant/insignificant variables distinction with real tick by tick stock data.
- We evaluate our proposed model through prediction accuracy measure combined SAX and Jaro-Winkler distance for easily comparing singular points of the predicted graph and the actual graph.
- To generate stock price predicted automatically, we build up a new system based on big data processing open source tools such as Hadoop and R.

The remainder of this paper is organized as follows. Section 2 presents the background to understanding stock researches and introduces the characteristics of that data as big data. In Section 3, we describe target environment and define problem. Section 4 and 5 describe our new complex methodology and system architecture for handling overall processes. Section 6 presents our experiments for proving our proposal. In Section 7, we review various existing researches based on stock price forecasting. Finally, Section 8 concludes the paper.

2. Background and Stock data

In this section, we first introduce various issues for finance especially, stock prediction and then, we identify big stock data collected by Koscom.

2.1. Background for Stock research

There are major issues related to financial analysis, such as cloud computing, stock prediction, and data security. Firstly, the cloud computing has taken center stage in financial field [9, 11, 12, 30]. Especially, Chang [9] suggested the Heston model based on cloud computing to solve constraints of the desktop, which calculates asset prices, volatility, and so on in stocks.

As second issue, a lot of related papers have already been published and stock prices are still being predicted using a variety of methods such as machine learning, feature selection in the stock prediction [20, 37, 38]. However, these methods are suitable for predicting the closing price with a low liquidity because the number of the daily data with closing price is not large. So, recently, large scale data processing issue is being actively discussed to overcome the previous limitation [5, 12]. In [5], they use twitter data (9,853,498 tweets) for stock prediction and Chang et al. [12] have developed Organisational sustainability modelling (OSM) that can process thousands of data quickly in finance.

Finally, data security in finance is a subject that is constantly being discussed [36]. They have developed a Cloud Computing Adoption Framework (CCAF) for securing cloud data, it can protect data in real-time and support various functions such as intrusion prevention and convergent encryption.

In this paper, we deal with stock prediction in real data and big data processing as large scale data among issues mentioned above.

2.2. Historical stock data as big data

We were able to obtain a big dataset of historical stock data, as detailed in Table 1. Because the stock data indicate transaction amounts and volume on stock trading from Korea Composite Stock Price Index (KOSPI), we call it tick by tick data and they were collected by the Koscom Corporation from August 2014 to October 2014. The scale of the data collected is 10$\sim$15 GB per month and the number of the data is about 6.5 million per month; the size of total dataset is 50 GB and about twenty million.

However, because there are a lot of unnecessary variables such as several codes and numbers among fifty one variables in Table 1, we cannot use them directly. Therefore, this paper considers only the price and the amount of selling and buying to affect the stock prices and be used mainly from several studies, as shown in Table 2 [17, 38]. This table is composed of the date, time, item code, type, trade price, trade amount, opening price, high price and low price. Moreover, there are two kinds of country type: domestic (code: 00) and foreign, there are two kinds of investor type: individual (code: 8000) and institutional and there are two kinds of

Table 1: Raw stock data as tick by tick

	Total fifty one variables →					
	TRADE_DATE	**ISIN_CODE**	**TRD_TM**	**TRD_PRC**	**TRDVOL**	**...**
6.5 million per month ↓	20141001	KR7005380001	090000001	190000	10	...
	20141001	KR7005380001	090000001	190000	3	...
	20141001	KR7005380001	090000838	190000	27	...
	20141001	KR7005380001	090000984	190000	40	...
	...	...	...	...	...	...

Table 2: Example of stock raw data

Feature	**Value**
Date (yyyymmdd)	20140813
Time (hhmmssmmm)	090024000
ISIN code	KR7005380001
Country type	00
Investor type	8000
Trading type	BID
Trade price (won)	77,500
Trade amount	37
Opening price (won)	78,900
High price (won)	78,900
Low price (won)	76,600

trading type: buying (code: BID) and selling (code: ASK), stock price is totally sum of forty features.

3. Target environment and Problem definition

In this section, we firstly introduce a target environment to predict stock price and then, present the importance and problem of data selection in the past.

3.1. Target environment

Stock trading service aims to determine whether a price of specific item goes up and down and to take maximum profit through buying and selling at a reasonable prices. In this time, when we predict the stock price graph of the following day, it is necessary to judge the current situation such as sharply falling trading

price and gradually increasing trading amounts. Figure 2 describes trading price and several features to comprise it in five-minute intervals during one month (August 2014) in HYUNDAI motor company. For example, at first sight, the trading prices between two days (August 6 and 29, 2014) are similar in Figure 2 (a). However, it shows that they are different in Figure 2 (b)∼(f). Also, Figure 2 (g) does not seem to greatly affect the trading price.

Therefore, to predict the stock price of the following day, as shown in Figure 3, we need to use historical big stock data generated by transaction for retrieving similar situation with the current, unlike [1, 19, 38] that predict the one day future closing price of individual stocks using daily stock prices composed of small data. In other words, because the trading price consists of several features such as trading amount, high price and low price, this study is to find the most similar pattern to a combination of features among historical stock data and to predict the stock price using them. And, to determine time range to be predicted in this paper, it is necessary to define the stock price prediction, it can be defined as Definition 3.1.

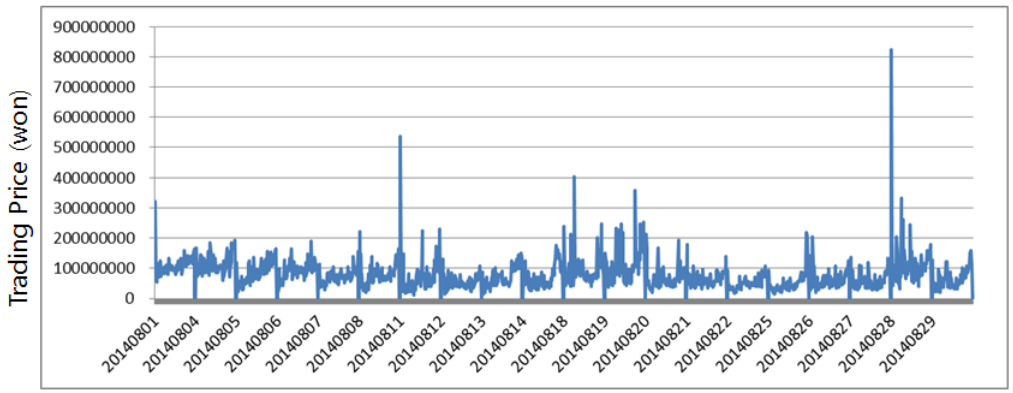
Definition 3.1: Stock price prediction

The stock price prediction is to forecast future stock price in the same period and item using similar patterns to the current pattern among historical stock data.

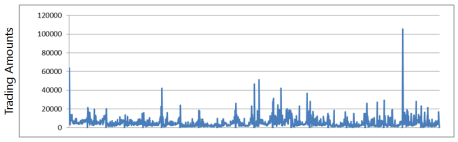
3.2. Problem in use of historical data

As previously discussed, because the stock data consists of several features, if historical data as input data for training in a prediction method without considering relation of them are just used, it might cause large residual between real and predicted data. For example, as shown in Figure 4, given a current pattern with one day size in trading price of a item, if historical stock data for a certain period of time are used when stock pattern of following day is generated, a gap between real and predicted stock graph is increased due to outliers such as sharply falling situation, these outliers also have been studied by several papers [2, 13].

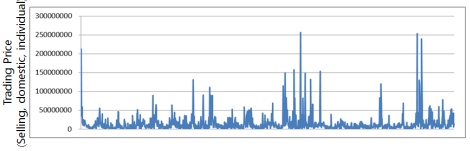
In this paper, the problem mentioned above can be defined as Definition 3.2



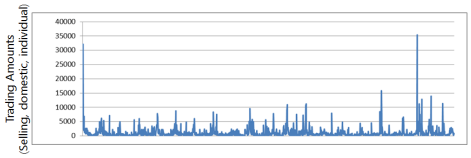
(a) Trading price for August 2014 in HYUNDAI motor company



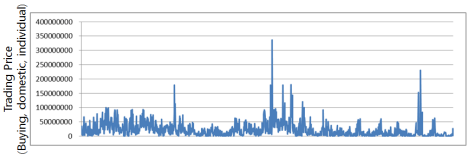
(b) Trading amounts for August 2014 in HYUNDAI motor company



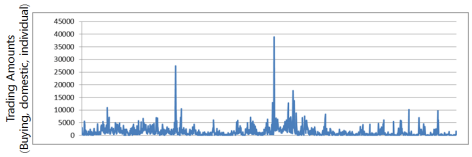
(c) Trading price at domestic individual selling for August 2014 in HYUNDAI motor company



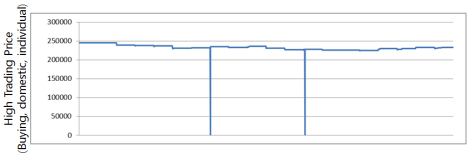
(d) Trading amounts at domestic individual selling for August 2014 in HYUNDAI motor company



(e) Trading price at domestic individual buying for August 2014 in HYUNDAI motor company



(f) Trading amounts at domestic individual buying for August 2014 in HYUNDAI motor company



(g) High trading price at domestic individual buying for August 2014 in HYUNDAI motor company

Figure 2: Trading price consists of various kinds of features.

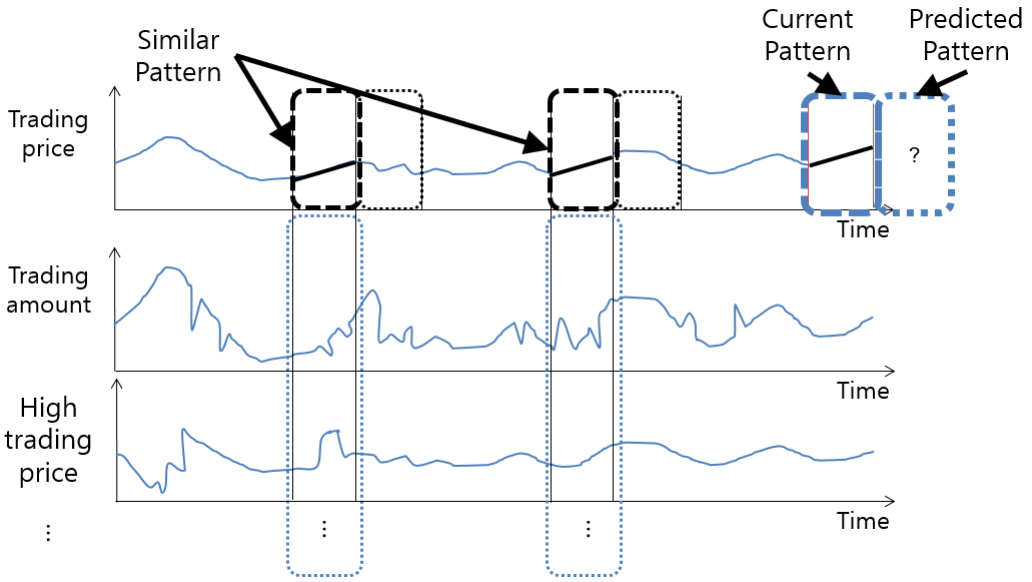


Figure 3: Predicted pattern will be generated from similar patterns of historical stock data.

Definition 3.2: Selection criteria absence of input data for prediction

Among historical stock data with various stock price pattern graph, to retrieve specific criteria for selecting optimal historical data that can increase

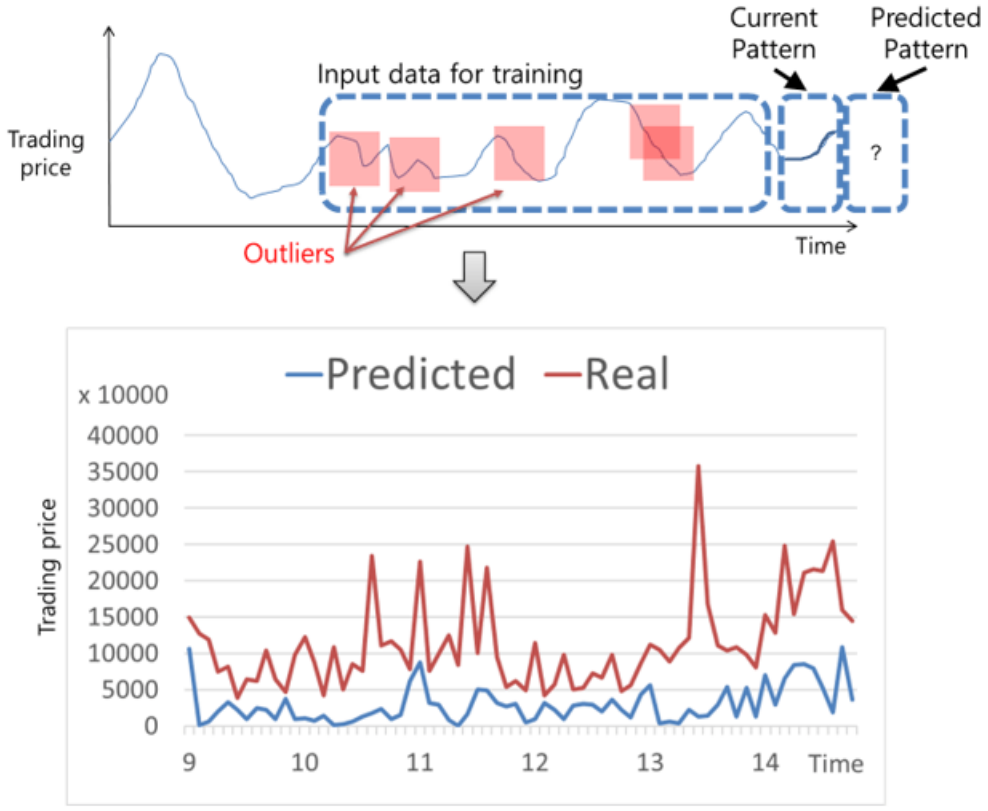


Figure 4: Example of problem in the selection of historical data.

the prediction accuracy is defined as the selection criteria absence of input data for prediction.

Definition 3.2 is also expressed by a formula as follows.

$$f^* = \underset{f}{\operatorname{argmin}} \sum^{i} \sum_{t_{i,j} \in T}^{j} L(S^{d-\gamma}_{t_{i-\alpha,j-\beta}}, forecast(s^{d}_{t_{i,j}})) \tag{1}$$

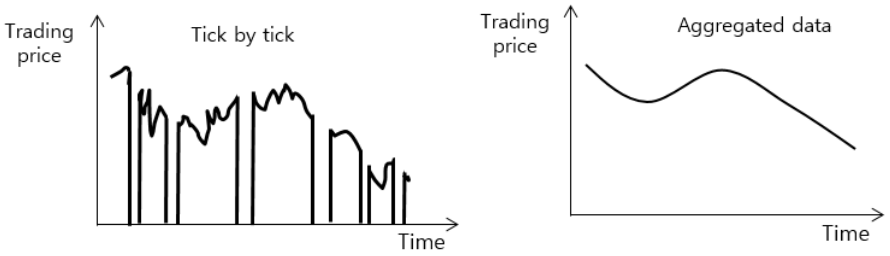
where f^* is to find suitable historical data with the least loss between real and predicted data, $\underset{f}{\operatorname{argmin}} f(x)$ means a function that finds x value to minimum $f(x)$. And, as $S^{d}_{t_{i,j}}$ is historical dataset, it can be expressed $\{S^{d}_{t_{i,j}}, S^{d-1}_{t_{i,j}}, S^{d-2}_{t_{i,j}}, \ldots, S^{d-n}_{t_{i,j}}\}$, t_i is hour, t_j is minute and d is day. $forecast$ function means a prediction method to use historical data as input data and L function is a loss method between real and predicted data.

4. Outline of proposed model

In this section, we describe the overall process, from data preprocessing for making continuous data, retrieving similar pattern data and selecting input data to the generation of prediction model from the perspective of data analysis and processing.

4.1. Data Preprocessing: Aggregation of stock data

We have tick by tick data received from Koscom. Because the data is generated per a transaction during a very short time, the trading price of tick by tick is zero at the time if the transaction is not carried out, as shown in Figure 5 (a) and eventually, the data is a discrete data not continuous data, it is difficult for the data to predict stock price. In other words, it is necessary to transform discrete data into continuous data without zero rather than discrete data as raw data considering zero for easily predicting the stock price. Consequently, we generate aggregated data at five-minute intervals to revise a continuous flow of data in Figure 5 (b).



(a) Trading price per transaction. (b) Trading price after aggregation.

Figure 5: The need for aggregation in raw stock data

4.2. Stock Pattern generation with Sliding Window

Because a stock graph shows a similar pattern aperiodically, we should find the dataset with similar patterns in big historical stock data. To do that, above all, it is necessary to make patterns from aggregated data. Figure 6 shows the processes of patterning the aggregated data. The length of a pattern is one day and patterns are generated at five-minute intervals, e.g., by the sliding window method, for pattern matching analysis using various patterns. The number of patterns for one hour will be twelve.

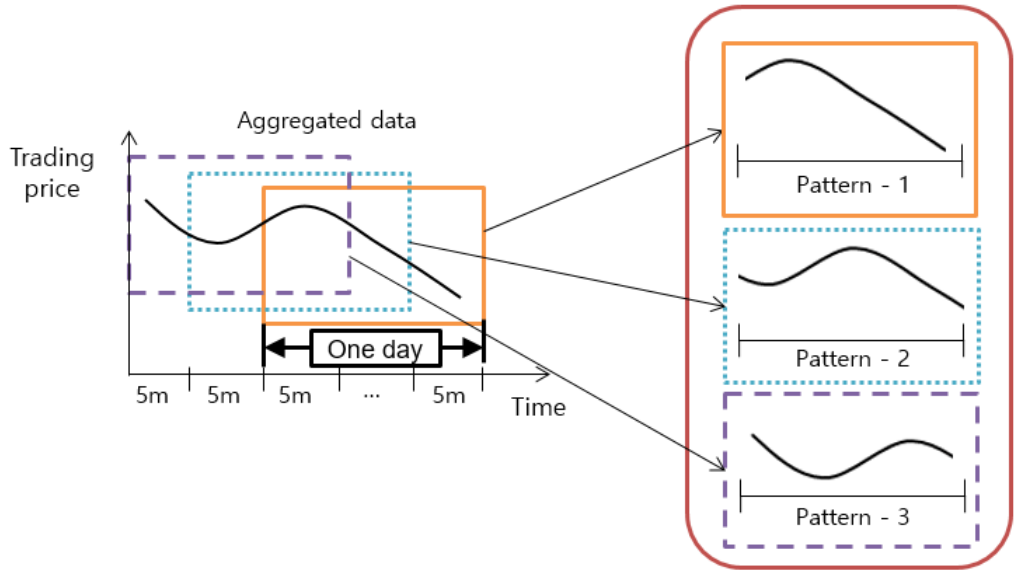


Figure 6: Method of patterning the aggregated data.

4.3. Pattern Retrieve using Dynamic Time Warping

Figure 7 shows seven similar patterns (dotted lines) and one current pattern (solid line) using Dynamic Time Warping in the graph of real stock price. The similar patterns can be found by comparing historical patterns and the current pattern. There are various methods for pattern matching such as Euclidean distance, Dynamic Time Warping (DTW)([4]), Edit Distance with Real Penalty (ERP)([14]), Longest Common Subsequence (LCSS)([39]) and Edit Distance on Real Sequence (EDR)([15]). Although we use a hierarchical clustering algorithm based on Euclidean distance by reason of finding similar patterns quickly and simultaneously in previous paper ([23]), because the Euclidean distance method does not accurately identify trading price trends due to a limitation that i^{th} point in one sequence should be calculated with the i^{th} point in the other, we find similar patterns using DTW method that accurately identify trading price trends than anything else ([3, 16]).

Figure 8 depicts the difference of Euclidean distance and DTW. Whereas a i^{th} point in one graph indicates the i^{th} point as same location in the other at Euclidean distance, a i^{th} point in one graph connects several points in the other at DTW. As an extreme example, given a sine curve and a cosine curve, when calculating at Euclidean distance, as the distance between two curves is big, two curve are likely different patterns. On the other hand, if using DTW, two curve are likely similar

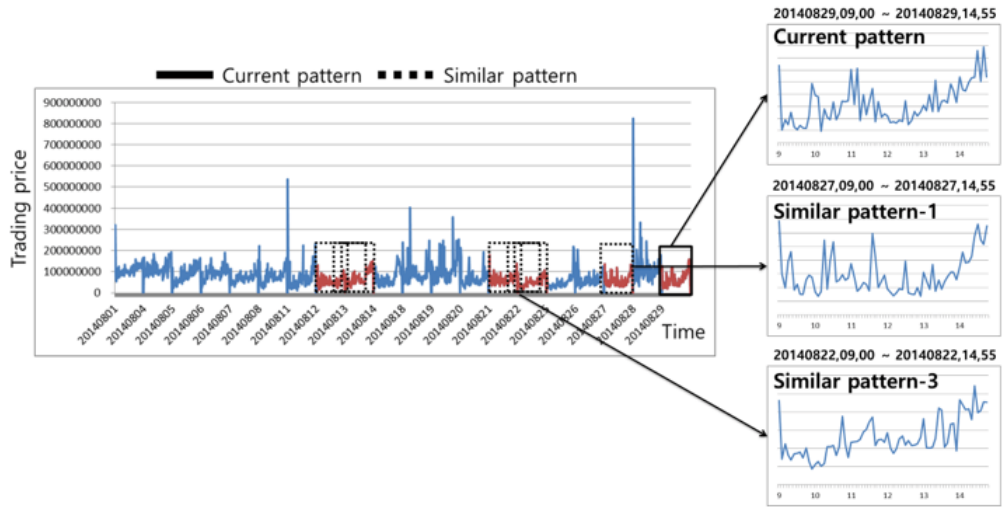
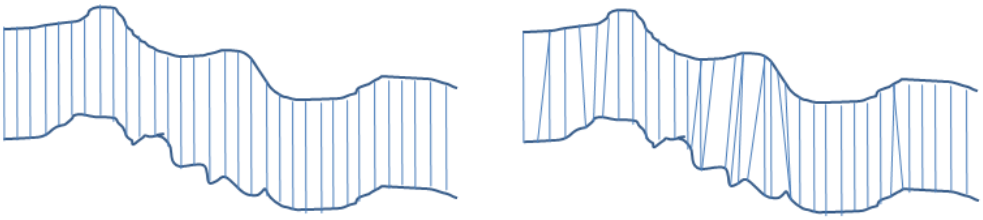


Figure 7: Similar stock patterns.



(a) Comparison of two data according to Euclidean

(b) Comparison of two data according to Dynamic Time Warping

Figure 8: Difference of Euclidean and Dynamic Time Warping

patterns. By doing this, it would be to find a pattern having a similar situation for the current pattern.

4.4. Feature Selection based on Stepwise Regression Analysis

After finding similar historical patterns with current trading price pattern, it is necessary to decide determinants which are most influenced by the trading price and retrieve optimal historical pattern dataset using them. While this seems same pattern between current and historical trading price, the relationship of the determinants may be different. For example, if current trading price is 10,000 won due to selling by institutional investors and historical trading price was 10,000 won due to selling by individual investors, they cannot be seen the same.

In this paper, for finding optimal historical pattern dataset, we choose main determinants using stepwise regression in both current and historical patterns, this is called "Feature selection". Before creating each regression model, it must be normalized because the units of total forty determinants are different. Because whereas the unit of the amount is the number, the unit of the price is won, dollar and yen, the units of the determinants are made into one for making it possible to compare determinants. Although there are various data transformation techniques such as z-transform, log transform and re-scaling range to [0,1], we use re-scaling range in R because it was often used for continuous data, as shown below, and two Table 3 and 4 show raw and transformed data.

x_i `<-(`x_i`-min(`x_i`) /(max(`x_i`)-min(`x_i`))`

Table 3: Before data transformation

	Total forty variables →				
	TRD_PRC	**TRDVOL**	**HIGH_PRICE in ASK, Individual and Domestic**	**LOW_PRICE in BID, Institutional and Foreign**	**...**
One thousand five hundreds per one month ↓	3509500	9	222000	218500	...
	4388000	147	222000	218500	...
	1533000	1221	218500	217000	...
	...	...	...	...	...

Given normalized forty determinants, we create regression model using them. In this work, we consider the trading price as a dependent variable and forty determinants as independent variables in the regression analysis, which is provided as two function in R as shown below. Above all, we use *lm* function to fit a linear model. The y is dependent variable and x_1 to x_{40} are independent variables.

`fit <- lm(`$y \sim x_1+x_2+x_3+...+x_{40}$`, data=stock_data)`

After fitting, we also use step function in R for determining the final independent variables, the first factor represents the linear model and the second factor determines the direction of the stepwise process combining forward and backward.

```
bidirectional <- step(fit, direction="both")
```

Table 4: After data transformation

	Total forty variables →				
	TRD_PRC	**TRDVOL**	**HIGH_PRICE in ASK, Individual and Domestic**	**LOW_PRICE in BID, Institutional and Foreign**	**...**
One thousand five hundreds per one month ↓	0.234579665	0.239762641	0.785714286	0.770877944	...
	0.487807624	0.090879625	0.785714286	0.770877944	...
	0.184445404	0.748927895	0.781512605	0.773019272	...
	...	...	...	...	...

Table 5: This table shows result of stepwise regression in real stock data of Hyundai Motor Company.

	ASK				BID			
	Domestic		Foreign		Domestic		Foreign	
	Indiv.	Insti.	Indiv.	Insti.	Indiv.	Insti.	Indiv.	Insti.
Trading Price	O	X	O	X	O	O	O	O
Trading Volume	O	X	X	O	O	O	O	X
High Price	X	X	X	X	X	X	X	X
Low Price	X	X	O	O	X	O	O	X
Opening Price	X	X	X	X	X	X	X	X

The procedures is organized as follows.

- Repeatedly add and remove a variable among all variables, then conduct regression analysis with remainder.
- Select final variable association with the highest value of R-Square as explanatory power of regression model.

In current pattern of Hyundai Motor Company, a total of fifteen variables are remained after applying stepwise regression, as can be seen in Table 5.

4.5. Predicted Stock data generation using Artificial Neural Network

By using comparison of important determinants in both current and historical patterns, we select optimal historical dataset and then, they will be used for input

data of Artificial Neural Network. First, we calculate and compare leverage of determinants in each regression model using lm.beta function provided by R. In this time, it is determined by the number of same elements equal to or greater than the threshold value of leverage, we can check the results in Table 6.

And, Table 6 shows how many matching numbers between current and historical patterns to judge the similar pattern. Because the matching number in the second historical pattern is just two whereas the matching number in the first historical pattern is four, we select first historical pattern as optimal historical dataset.

After selecting historical dataset, to generate predicted stock data, we use an artificial neural network algorithm because the algorithm is the most widely used in stock price forecasts ([28, 24, 8]) and has a high predictive power as an advantage through learning by iterative adjustment. Here, optimal historical dataset will be used for training data as input data in artificial neural network. As shown below R code, an ANN model is created with one dependent and four independents using neuralnet package ([18]).

```
neural <- neuralnet(TRD_PRC ~ BID_DOM_INDIV_TRD_PRC
+ BID_DOM_INSTI_TRD_PRC + BID_FOR_INDIV_TRD_PRC
+ BID_FOR_INSTI_TRD_PRC, data = training_data, hidden=3)
```

Because the units of input data are converted to [0,1] transform, the result of ANN must be also converted to previous unit. As shown below R code, we generate predicted stock data using test data and ANN model based on optimal historical data. Then, it converts the unit of the predicted data to the original unit.

```
neural_results <- compute(neural, test_data)
final_ANN_results <- neural_results$net.result*(max(x)-min(x))+min(x)
```

4.6. Prediction accuracy measure combined SAX and Jaro-Winkler

There are various measures such as Beta (β), Standard Error, MSE, R-squared value, MAPE and RMSE for checking prediction accuracy in finance [10]. However. in this paper, we use a combined SAX and Jaro-Winkler as new prediction accuracy measure to approximately find singular points of patterns because the pattern does not know when and where it occurs.

This measure consists of two methods, SAX and Jaro-Winkler distance. As the SAX is to transform time-series data into approximated string, it consists of three steps. First, it transforms the time-series data into the normalized time-series data to pick equal-sized areas, as shown Figure 10. Second, it transforms the normalized data into the Piece-wise Aggregate Approximation (PAA) [42]

Table 6: Selection of optimal historical dataset

(a) Calculation of determinants leverage

Leverage of determinants in current pattern		**Leverage of determinants in historical pattern - 1**		**Leverage of determinants in historical pattern - 2**	
Determinant	Leverage	Determinant	Leverage	Determinant	Leverage
BID_DOM_INSTI_TRD_PRC	0.340	BID_DOM_INSTI_TRD_PRC	0.318	BID_DOM_INSTI_TRD_PRC	0.000
BID_DOM_INSTI_TRDVOL	0.000	BID_DOM_INSTI_HIGH_PRICE	-0.000	BID_DOM_INSTI_TRDVOL	0.000
BID_DOM_INSTI_LOW	0.000	BID_DOM_INSTI_LOW	-0.000	BID_DOM_INSTI_HIGH_PRICE	-0.000
ASK_FOR_INSTI_TRDVOL	0.000	ASK_FOR_INSTI_TRD_PRC	0.549	ASK_FOR_INSTI_TRD_PRC	0.000
ASK_FOR_INSTI_LOW	-0.000	ASK_FOR_INSTI_TRDVOL	-0.000	ASK_FOR_INSTI_TRDVOL	-0.000
BID_FOR_INDIV_TRD_PRC	0.011	BID_FOR_INDIV_TRD_PRC	0.038	ASK_FOR_INSTI_LOW	-0.000
BID_FOR_INDIV_TRDVOL	-0.000	BID_FOR_INDIV_LOW	0.000	BID_FOR_INDIV_TRD_PRC	0.005
BID_FOR_INDIV_LOW	0.000	BID_DOM_INDIV_TRD_PRC	0.585	BID_FOR_INDIV_HIGH	-0.000
BID_DOM_INDIV_TRD_PRC	0.794	ASK_DOM_INDIV_TRD_PRC	0.687	BID_DOM_INDIV_TRD_PRC	0.000
BID_DOM_INDIV_TRDVOL	-0.000	ASK_DOM_INDIV_TRDVOL	-0.000	ASK_DOM_INDIV_TRD_PRC	0.000
ASK_DOM_INDIV_TRD_PRC	0.000	ASK_FOR_INDIV_TRD_PRC	-0.000	ASK_DOM_INDIV_TRDVOL	-0.000
ASK_DOM_INDIV_TRDVOL	-0.000	ASK_FOR_INDIV_TRDVOL	-0.000	BID_FOR_INSTI_TRD_PRC	0.471
ASK_FOR_INDIV_TRD_PRC	0.000	ASK_FOR_INDIV_HIGH	0.000	**Total number of determinants**	**12**
ASK_FOR_INDIV_LOW	-0.000	ASK_FOR_INDIV_LOW	-0.000		
BID_FOR_INSTI_TRD_PRC	0.389	BID_FOR_INSTI_TRD_PRC	0.713		
Total number of determinants	**15**	**Total number of determinants**	**15**		

(b) Final leverage applied by threshold

Leverage of determinants in current pattern		**Leverage of determinants in historical pattern - 1**		**Leverage of determinants in historical pattern - 2**	
Determinant	Leverage	Determinant	Leverage	Determinant	Leverage
BID_DOM_INSTI_TRD_PRC	0.340	BID_DOM_INSTI_TRD_PRC	0.318	ASK_FOR_INSTI_TRD_PRC	0.549
BID_FOR_INDIV_TRD_PRC	0.011	BID_FOR_INDIV_TRD_PRC	0.038	BID_FOR_INDIV_TRD_PRC	0.005
BID_DOM_INDIV_TRD_PRC	0.794	BID_DOM_INDIV_TRD_PRC	0.585	ASK_DOM_INDIV_TRD_PRC	0.687
BID_FOR_INSTI_TRD_PRC	0.389	BID_FOR_INSTI_TRD_PRC	0.713	BID_FOR_INSTI_TRD_PRC	0.471
Total number of determinants	**4**	**Matched number of determinants**	**4**	**Matched number of determinants**	**2**

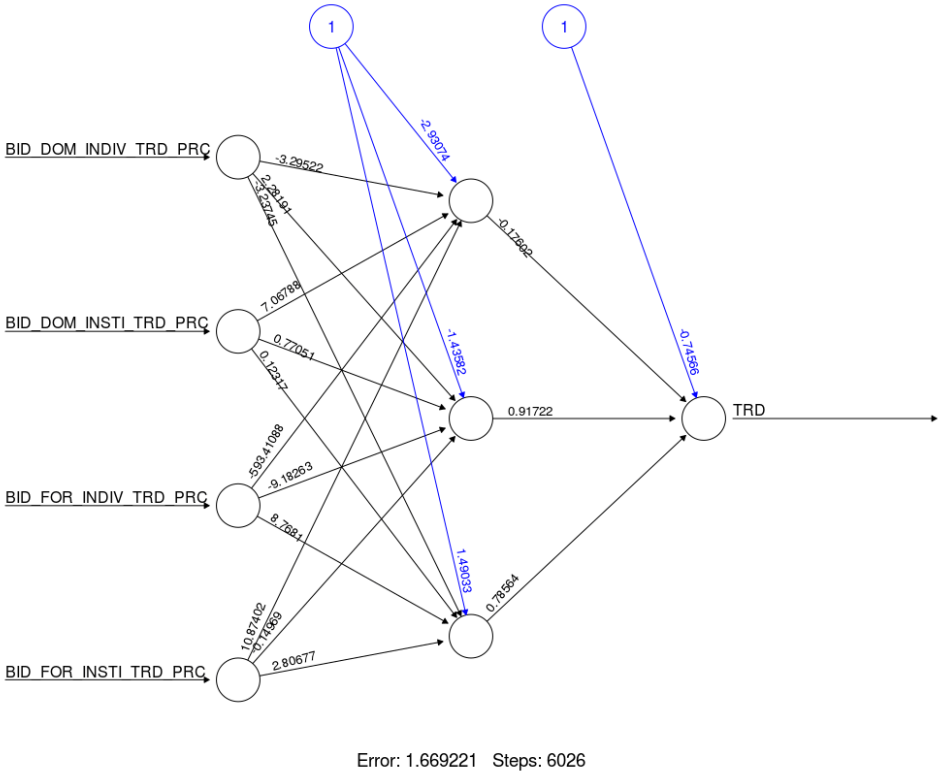


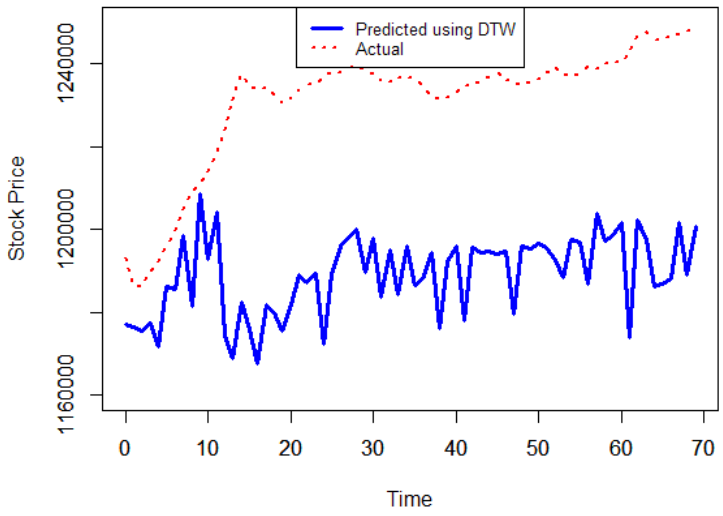
Figure 9: Graphical representation of ANN model with hidden layer 3

representation as shown Figure 11 and finally, it converts the PAA data into a string as follows.

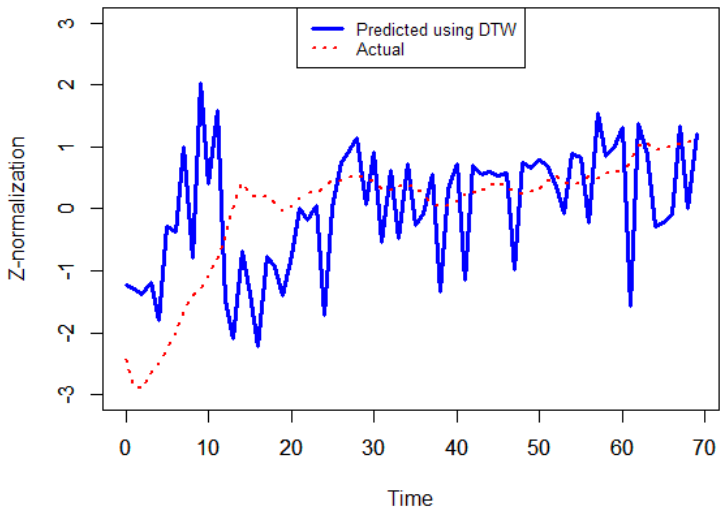
```
Predicted data: adabcccddddd
Real data: aaccddcdddde
```

Then, the Jaro-Winkler distance is to compare two transformed strings and calculate similarity(accuracy). As this distance is two complexed distance measures, Jaro distance is expressed by a formula as follows.

$$distance_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{2}$$

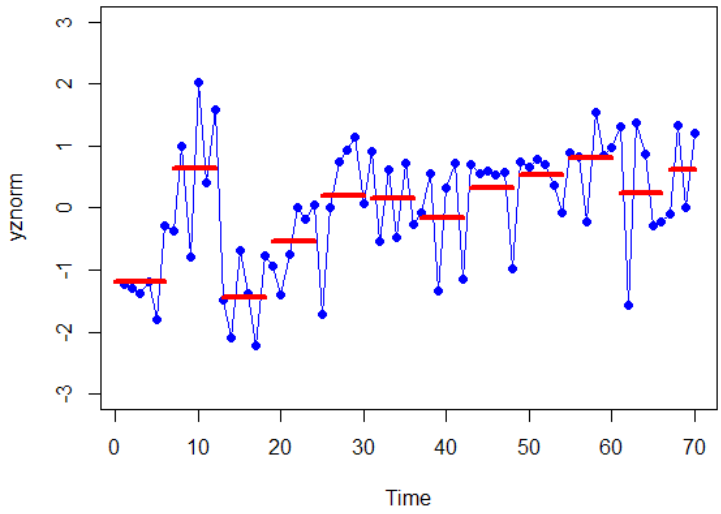


(a) Two predicted graph and actual graph

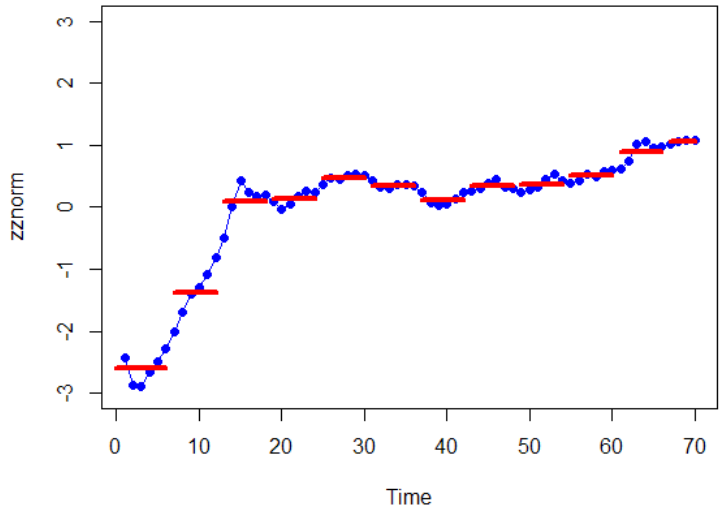


(b) Two transformed graphs

Figure 10: Z-normalization for comparison



(a) Transform predicted data into PAA



(b) Transform actual data into PAA

Figure 11: PAA transformation

where s_1 and s_2 are strings to compare, m is the number of matching characters and t is half the number of transpositions. Eventually, the Jaro-Winkler distance is expressed by a formula as follows.

$$distance_{jw} = distance_j + l * p(1 - distance_j) \tag{3}$$

where l is the length of common prefix at the start of the string up to a maximum of 4 characters and as p is a constant scaling factor, the standard value for this constant is p=0.1. Through this distance, we have get the similarity from next formula.

$$similarity = 1 - distance_{jw} \tag{4}$$

According to above formulas, the similarity of the two strings mentioned above is 88.18%.

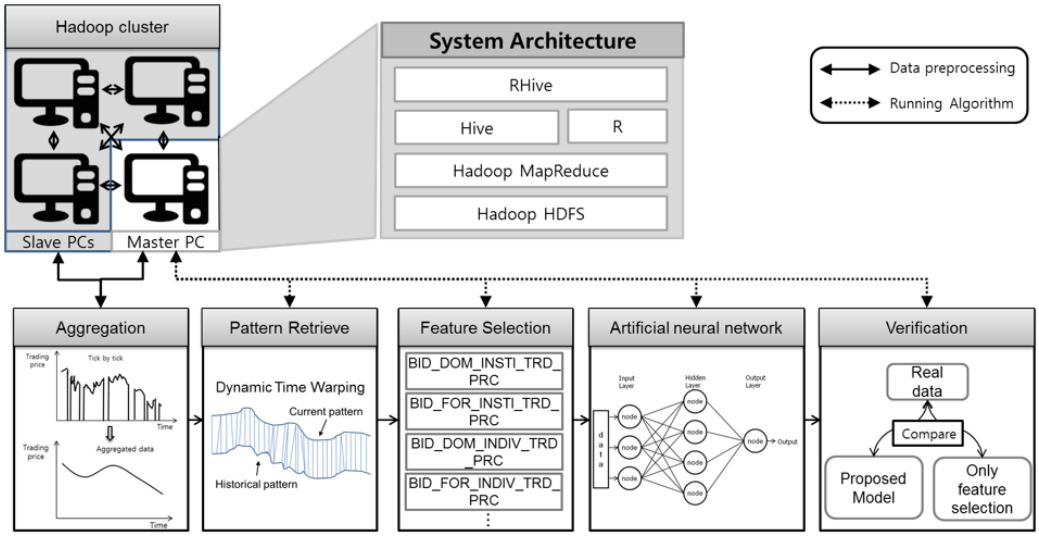


Figure 12: System Architecture

5. System architecture for stock price prediction

This section describes the series of operations that were implemented when generating the final artificial neural network model. All the processes were conducted on cluster composed of four connected computers (one master and three slaves) with Hadoop and RHive installed.

5.1. A series of operations for generating predicted stock data

We suggest the following steps for generating a prediction model on bigdata processing and analysis tools, as shown in Figure 12.

Step 1 (Stock data aggregation and pattern generation as data preprocessing): We store the stock data for three months provided by the KOSCOM in Hadoop Distributed File Systems (HDFSs) of the Hadoop based cluster. Because we cannot manually modify the source code of MapReduce for extracting exactly desired data from each HDFS of Hadoop cluster, we use RHive tool to provide HiveQL that assists to search the desired data such as select query of RDBMS. The HiveQL queries such as 'CREATE', 'LOAD' and 'SELECT' are as follows.

```
rhive.query(CREATE TABLE STOCK_PREDICTION
(TRADE_DATE STRING, BLKTRD_TP_CD STRING,
REGUL_OFFHR_TP_CD STRING, ISIN_CODE STRING, JONG_INDEX INT,
TRD_NO INT, TRD_PRC FLOAT, TRDVOL INT, TRD_TP_CD STRING,
TRD_DD STRING, TRD_TM STRING, NBMM_TRD_PRC FLOAT,
FUTRMM_TRD_PRC FLOAT, BID_MBR_NO STRING,
BIDORD_TP_CD STRING, BID_TRSTK_STAT_ID STRING,
BID_TRSTK_TRD_METHD_CD STRING, BID_ASK_TP_CD STRING,
BID_TRST_PRINC_TP_CD STRING, BID_TRSTCOM_NO STRING,
BID_PT_TP_CD STRING, BID_INVST_TP_CD STRING,
BID_FORNINVST_TP_CD STRING, BIDORD_ACPT_NO INT,
ASK_MBR_NO STRING, ASKORD_TP_CD STRING,
ASK_TRSTK_STAT_ID STRING, ASK_TRSTK_TRD_METHD_CD STRING,
ASK_ASK_TP_CD STRING, ASK_TRST_PRINC_TP_CD STRING,
ASK_TRSTCOM_NO STRING, ASK_PT_TP_CD STRING,
ASK_INVST_TP_CD STRING, ASK_FORNINVST_TP_CD STRING,
ASKORD_ACPT_NO INT, OPEN_PRICE FLOAT, HIGH_PRICE FLOAT,
LOW_PRICE FLOAT, LST_PRC FLOAT, ACC_TRDVOL INT,
ACC_AMT FLOAT, LST_ASKBID_TP_CD STRING, LP_HD_QTY INT,
DATA_TYPE INT, MSG_SEQ INT, BID_PROGM_ORD_DECL_TP_CD STRING,
ASK_PROGM_ORD_DECL_TP_CD STRING, BRD_ID STRING,
SESSION_ID STRING, DYNMC_UPLMTPRC INT, DYNMC_LWLMTPRC INT)
PARTITIONED BY (TRADE_DATE STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',
LINES TERMINATED BY '\n')

rhive.query(LOAD DATA LOCAL INPATH
'KOSPI/201408/*.txt' OVERWRITE INTO
TABLE STOCK_PREDICTION PARTITION(DATE= '201408')
```

```
rhive.query(SELECT * FROM STOCK_PREDICTION
WHERE ISIN_CODE='KR7005380001' AND TRADE_DATE LIKE '201408%'
ORDER BY TRADE_DATE ASC)
```

After extracting the data, they are aggregated at five minutes intervals by using R because of tick by tick data. Then, patterns are generated from them because of concatenation of similar patterns in R of the master computer. The size of a pattern is one day and generation unit was five minutes intervals.

Step 2 (Pattern selection with dynamic time warping): To retrieve similar patterns with current pattern, we use dtw as dynamic time warping in dist function that measure a distance in R, which has the powerful two advantages that it does not unnecessary comparison operation irrelevant to the current pattern rather than hierarchical cluster supporting euclidean distance and it can more accurately detect similar patterns.

Algorithm 1: Algorithm for pattern selection

input : List of aggregated patterns P_A and a current pattern P_C
output: List of similar patterns P_S after 'DTW' method is applied

1 Initialize *last* to size of aggregated patterns P_A;
2 **for** $i = 1 \rightarrow last$ **do**
3 Calculate distance D_i based on Dynamic Time Warping between i^{th} aggregated pattern P_{A_i} and current pattern P_C;
4 Add D_i in array of integer V_{DTW};
5 Sort V_{DTW} in ascending order;
6 Extract Top-ten patterns of sorted V_{DTW} to P_S as the most similar patterns;
7 return P_S;

Algorithm 1 describes to find Top-ten similar patterns. After inserting the current pattern into aggregated patterns as historical dataset, we calculate dtw-based distance between current pattern and a pattern generated by sliding window method (Line 2 $\sim$ 4). Real codes in R is as follows. Then, we select top ten patterns with the smallest difference (Line 5 $\sim$ 6).

```
P_A.dist <- dist(P_A, method='DTW')
```

Step 3 (Feature selection using stepwise regression): Given a current pattern of stock price, insignificant variables in all variables composed of the price are removed.

Algorithm 2: Algorithm for feature selection of current pattern in stepwise regression

input : A current pattern P_C
output: List of remained variables excluding the insignificant variables *Var*

```
1 Transform variables of current pattern P_C to normalization NM_C;
2 Extract STD_C to remained variables Var using stepwise regression method;
3 Initialize len to size of remained variables Var;
4 for i = 1 → len do
5     if P-value of i^th Var is more than 0.05 then
6         break;
7     else
8         flag = true;
9 return Var;
```

Algorithm 2 describes the steps for feature selection of current pattern using stepwise regression. Firstly, before stepwise regression, all variables are normalized due to the difference of the units of total forty variables (Line 2). Then, insignificant variables among variables of current pattern are removed using stepwise regression (Line 3). Remained variables with p value below a specified threshold are judged as significant variables (Line 5 $\sim$ 8). Real codes in R is as follows.

Var_{final} $\leftarrow$ `step(`$Var_{initial}$`, direction='both')`

After removing insignificant variables in current pattern, optimal historical dataset should be retrieved using them. To do that, it is necessary to calculate and compare leverage of determinants in each regression model with current and historical(=similar) patterns. If a historical pattern has the number of same elements equal to or greater than the threshold value of leverage compared with current pattern, we decide optimal dataset.

Algorithm 3 describes comparison of leverage between current and historical patterns. First, beta values as leverage of the regression model are calculated using the remained variables generated from Algorithm 2 (Line 1). And, like Algorithm 2, all variables of each similar pattern are normalized to make regression models (Line 4). After creating each stepwise regression model of similar patterns, two beta values of the current pattern and a similar pattern are compared one by one (Line 5 $\sim$ 7).

Algorithm 3: Algorithm for optimal historical dataset selection using standardization

input : List of remained variables of current pattern Var and a list of similar patterns P_S

output: A list of optimal patterns P_O after beta comparison

1 Create Beta value $P_C.beta$ calculated from remained variables Var using beta function;
2 Initialize len to size of similar patterns P_S;
3 **for** $i = 0 \rightarrow len$ **do**
4 Transform variables of i^{th} similar pattern P_{S_i} to normalization NM_{S_i};
5 Create $stepmodel_i$ using stepwise regression with NM_{S_i};
6 Calculate Beta value $P_{S_i}.beta$ calculated from $stepmodel_i$;
7 Compare $P_C.beta$ and $P_{S_i}.beta$;
8 return P_O;

Step 4 (Predicted data generation on artificial neural network): To create predicted data, we use artificial neural network after feature selection. Algorithm 4 describes generation steps for predicted data using ANN method. Among input data, we prepare dependent and independent variables as training data with other time zone because we predict next day of current pattern (Line 1). Namely, given historical time ht of similar pattern, the time of the dependent variable is $ht+1$ and the time of the independent variable is ht. After independent and dependent variables are bound, we generated ANN based model using neuralnet function provided by R (Line 2). Then, independent variables at current time t in the model are input and predicted data are generated (Line 3). Real codes in R is as follows.

```
TR ← cbind(TR_dep, TR_indep)
colnames(TR) ← c('output','input')
ANN ← neuralnet(output∼input, training, hidden=3)
PRD ← compute(ANN, TEST_indep)
```

Step 5 (Verification): To verify our proposed model, we combine two methods such as SAX and Jaro-Winkler distance, the functions are also provided in R. The measure is computed from comparisons between real and predicted data. As shown below R code, it transforms two raw time-series data (predicted by DTW and real data) into z-normalized time-series data firstly. Then, after dividing them into twelve sections, each character is converted to five strings using

Algorithm 4: Algorithm for generation of predicted data

input : Total trading price TR_{dep} as training data, reminder variables TR_{indep} excluding total trading price as training data, and reminder variables $TEST_{indep}$ excluding total trading price as test data

output: A predicted trading price *PRD* generated by ANN

1 Bind dependent TR_{dep} and independent TR_{indep} variables;
2 Run artificial neural network *ANN* with bound variables;
3 Create predicted trading price *PRD* according to *ANN* with test variables $TEST_{indep}$;
4 return *PRD*;

series_to_string as string conversion function. Finally, a similarity between two strings is calculated using Jaro-Winkler distance function.

```
KR7005930003_pre_znorm = znorm(KR7005930003_pre)
KR7005930003_rea_znorm = znorm(KR7005930003_rea)
paa_size=12
s1_paa = paa(KR7005930003_20141027_pre_znorm, paa_size)
s2_paa = paa(KR7005930003_20141027_rea_znorm, paa_size)
str1 = series_to_string(s1_paa, 5)
str2 = series_to_string(s2_paa, 5)
stringsim(str1, str2, method='jw', p=0.1)
```

6. Evaluation

In this section, we describes the test data provided by the KOSCOM for three months and evaluate the accuracy of each stock item by computing SAX and Jaro-Winkler distance.

6.1. Dataset and test scenario

To prove our proposed model, we used real historical stock dataset composed of various items for three months from August 2014 to October 2014. To measure the prediction accuracy, we prepared three items (Hyundai motor, KIA motor and Samsung electronic) as companies representing Republic of Korea, its stock data for August 1, 2014 to October 26, 2014 as the training data, and its stock data for October 27 to 31, 2014 as the test data. As test scenario, first, two predicted stock data for one day were generated according to our proposed model and feature

selection. Then, we checked the prediction accuracy from Jaro-Winkler distance values comparing predicted and real stock data.

6.2. Evaluation of the prediction accuracy

We performed experiments to compute the accuracy of our proposed method. Figure 13 $\sim$ 15 describes comparison results of actual and predicted data by our proposed model in LG electronic, Samsung electronic and Hyundai motor company stock for five days (October 27 to October 31, 2014) as one week. The x-axis is the time at five minutes intervals and the y-axis is trading price called stock price according to the time.

In Figure 13 (a), we can directly know that our predicted graph and real graph are very similar, whereas Figure 13 (c) has a completely different trend. And, Figure 13 (b) and (d) show actually a big difference, but a little similar trend. Naturally, the Jaro-Winkler similarity is also similar to the result of the figure.

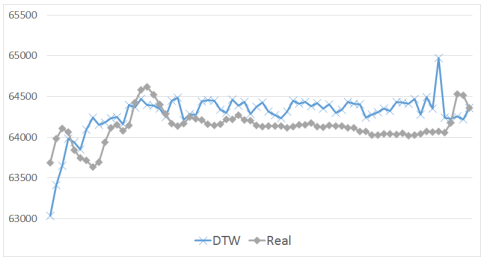
Figure 14 shows the stock data derived from the real and predicted data in Samsung electronic company. Like Figure 13, the more similar the trend, the higher the similarity in all graphs. Especially, although Figure 14 (d) is big difference between predicted and actual data, the similarity is high because the trend is very similar except opening time.

Lastly, Figure 15 depicts the stock data derived from the real and predicted data in Hyundai motor company. Seemingly, all figures have a high degree of similarity, due to the singular points of each figure, trends of predicted and actual data is different and the similarities are low compared to what we thought.

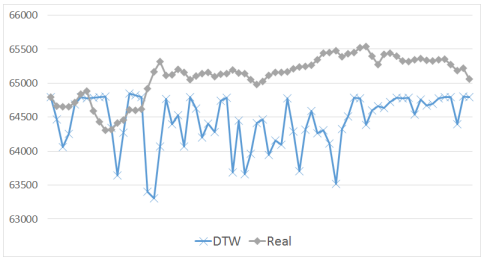
7. Related works

In this section, we introduce related works using various prediction methods such as artificial neural network, feature selection and text mining for stock price prediction. First, an artificial neural network, as the most widely used method from a few decades ago, was used by itself, and gradually attempted to combine with other techniques for a higher prediction accuracy. In [26], they have proposed buying and selling timing prediction system using economic indexes (foreign exchange rates) and technical indexes (vector curves) on the Tokyo Stock Exchange Prices Indexes. In another research, they have used echo state network as novel recurrent neural network to forecast the next closing price ([32]).

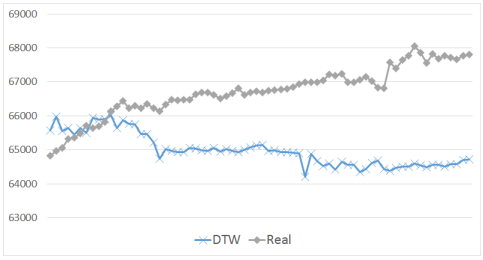
The following method is feature selection to select significant input attributes, as supporting the other methods and recently often used. [21] have proposed to combine support vector regression (SVR) with the self-organizing feature map



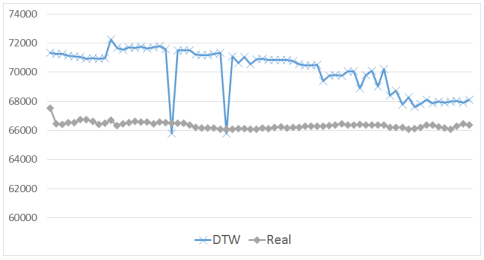
(a) 20141027



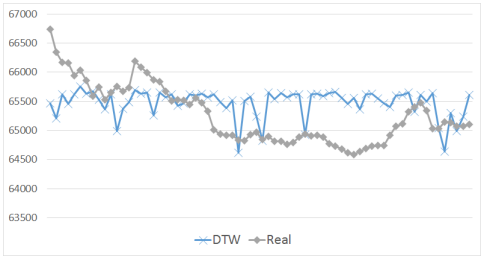
(b) 20141028



(c) 20141029



(d) 20141030



(e) 20141031

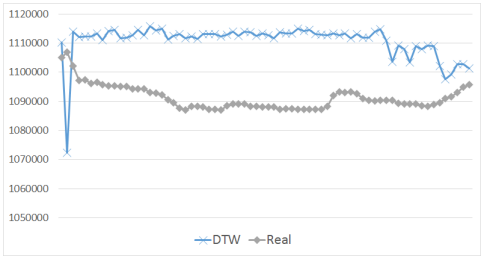
LG Electronic company	Jaro-Winkler Similarity
2016-10-27	0.7071429
2016-10-28	0.6666667
2016-10-29	0.4722222
2016-10-30	0.6527778
2016-10-31	0.4777778

(f) 20141031

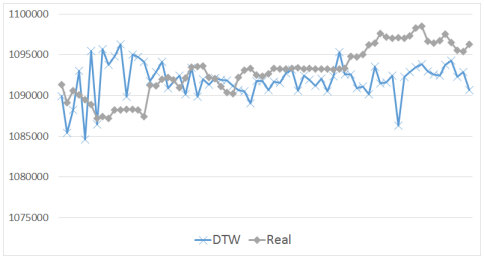
Figure 13: Comparison of real and predicted data in LG electronic company

(SOFM) technique and feature selection based on filtering for predicting the next day's price index on Taiwan index futures (FITX). In here, they have selected important features using r-squared value as input data of SVR. [29] have developed a prediction model based on support vector machine (SVM) with a hybrid feature selection method, which finds the original input features, on NASDAQ Index direction and unlike above paper, f-score has been used as a selection factor.

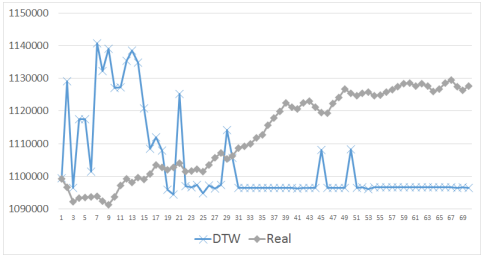
However, most of them have some limitations for short-term prediction. First, given all historical stock data as input data, because they have predicted the next closing price without removing outlier, there is a problem of a high error rate due to them. Second, it was insufficient to consider various factors although total



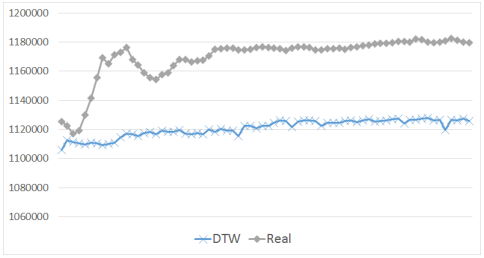
(a) 20141027



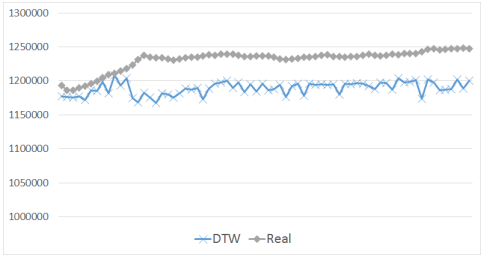
(b) 20141028



(c) 20141029



(d) 20141030

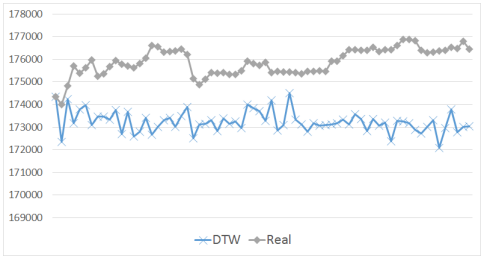


(e) 20141031

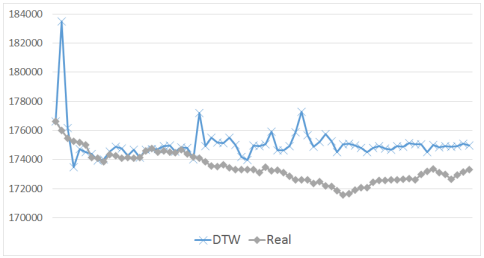
Samsung Electronic company	Jaro-Winkler Similarity
2016-10-27	0.4777778
2016-10-28	0.7071429
2016-10-29	0.3888889
2016-10-30	0.7071429
2016-10-31	0.8818182

(f) 20141031

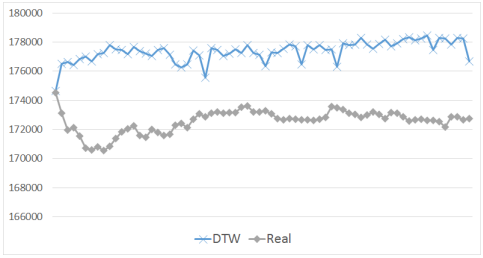
Figure 14: Comparison of real and predicted data in Samsung electronic company



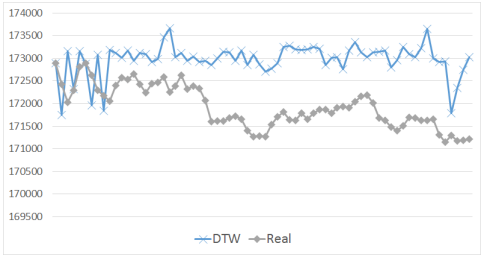
(a) 20141027



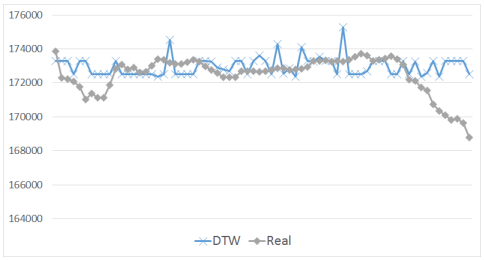
(b) 20141028



(c) 20141029



(d) 20141030



(e) 20141031

Hyundai motor company	Jaro-Winkler Similarity
2016-10-27	0.6031746
2016-10-28	0.6
2016-10-29	0.7222222
2016-10-30	0.4722222
2016-10-31	0.5555556

(f) 20141031

Figure 15: Comparison of real and predicted data in Hyundai motor company

completion price is determined by a variety of factors such as foreign purchase closing price, domestic selling completion amount. In other words, it is necessary to make a combination with some significant factors.

8. Conclusions

In this paper, we have determined that the stock prices sparsely show similar patterns and all of the variables are not a significant impact on the price. For short-term prediction, we proposed a novel method based on combination of dynamic time warping, stepwise regression and artificial neural network model to find similar historical datasets for each stock item and predict daily stock price using optimal significant variables through feature selection and comparison of leverage. Moreover, we dealt with the overall process using a big data processing framework composed of Hadoop, R and RHive. Finally, we demonstrated a prediction accuracy for three stock items using SAX and Jaro-Winkler distance.

References

[1] Adebiyi, A., Ayo, C., Adebiyi, M. O., Otokiti, S., 2012. Stock price prediction using neural network with hybridized market indicators. Journal of Emerging Trends in Computing and Information Sciences 3 (1), 1–9.

[2] Ané, T., Ureche-Rangau, L., Gambet, J.-B., Bouverot, J., 2008. Robust outlier detection for asia–pacific stock index returns. Journal of International Financial Markets, Institutions and Money 18 (4), 326–343.

[3] Banavas, G. N., Denham, S., Denham, M. J., et al., 2000. Fast nonlinear deterministic forecasting of segmented stock indices using pattern matching and embedding techniques. Computing in Economics and Finance 2000 64.

[4] Berndt, D. J., Clifford, J., 1994. Using dynamic time warping to find patterns in time series. In: KDD workshop. Vol. 10. Seattle, WA, pp. 359–370.

[5] Bollen, J., Mao, H., Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science 2 (1), 1–8.

[6] Bulkowski, T. N., 2011. Encyclopedia of chart patterns. Vol. 225. John Wiley & Sons.

[7] Cao, L., Tay, F. E., 2001. Financial forecasting using support vector machines. Neural Computing & Applications 10 (2), 184–192.

[8] 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. Computers & Operations Research 32 (10), 2499–2512.

[9] Chang, V., 2014. The business intelligence as a service in the cloud. Future Generation Computer Systems 37, 512–534.

[10] Chang, V., 2014. A proposed model to analyse risk and return for Cloud adoption. Lambert.

[11] Chang, V., Ramachandran, M., 2016. Towards achieving data security with the cloud computing adoption framework. IEEE Transactions on Services Computing 9 (1), 138–151.

[12] Chang, V., Walters, R. J., Wills, G. B., 2016. Organisational sustainability modellingan emerging service and analytics model for evaluating cloud computing adoption with two case studies. International Journal of Information Management 36 (1), 167–179.

[13] Charles, A., Darné, O., 2005. Outliers and garch models in financial data. Economics Letters 86 (3), 347–352.

[14] Chen, L., Ng, R., 2004. On the marriage of lp-norms and edit distance. In: Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. VLDB Endowment, pp. 792–803.

[15] Chen, L., Özsu, M. T., Oria, V., 2005. Robust and fast similarity search for moving object trajectories. In: Proceedings of the 2005 ACM SIGMOD international conference on Management of data. ACM, pp. 491–502.

[16] Coelho, M. S., 2012. Patterns in financial markets: Dynamic time warping. Ph.D. thesis, NSBE-UNL.

[17] de Oliveira, F. A., Nobre, C. N., Zarate, L. E., 2013. Applying artificial neural networks to prediction of stock price and improvement of the directional prediction index–case study of petr4, petrobras, brazil. Expert Systems with Applications 40 (18), 7596–7606.

[18] Fritsch, S., Guenther, F., Guenther, M. F., 2012. Package neuralnet. Training of Neural Network 1.

[19] Hsu, C.-M., 2013. A hybrid procedure with feature selection for resolving stock/futures price forecasting problems. Neural Computing and Applications 22 (3-4), 651–671.

[20] Huang, C.-F., Chang, B. R., Cheng, D.-W., Chang, C.-H., 2012. Feature selection and parameter optimization of a fuzzy-based stock selection model using genetic algorithms. International Journal of Fuzzy Systems 14 (1), 65–75.

[21] Huang, C.-L., Tsai, C.-Y., 2009. A hybrid sofm-svr with a filter-based feature selection for stock market forecasting. Expert Systems with Applications 36 (2), 1529–1539.

[22] Ince, H., Trafalis, T. B., 2007. Kernel principal component analysis and support vector machines for stock price prediction. IIE Transactions 39 (6), 629–637.

[23] Jeon, S., Hong, B., Lee, H., Kim, J., 2016. Stock price prediction based on stock big data and pattern graph analysis. In: Proceedings of the International Conference on Internet of Things and Big Data. pp. 223–231.

[24] Kim, K.-j., Han, I., 2000. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. Expert systems with Applications 19 (2), 125–132.

[25] Kim, Y., Jeong, S. R., Ghani, I., 2014. Text opinion mining to analyze news for stock market prediction. Int. J. Advance. Soft Comput. Appl 6 (1).

[26] Kimoto, T., Asakawa, K., Yoda, M., Takeoka, M., 1990. Stock market prediction system with modular neural networks. In: Neural Networks, 1990., 1990 IJCNN International Joint Conference on. IEEE, pp. 1–6.

[27] Kohara, K., Ishikawa, T., Fukuhara, Y., Nakamura, Y., 1997. Stock price prediction using prior knowledge and neural networks. Intelligent systems in accounting, finance and management 6 (1), 11–22.

[28] Kuo, R. J., Chen, C., Hwang, Y., 2001. An intelligent stock trading decision support system through integration of genetic algorithm based fuzzy neural

network and artificial neural network. Fuzzy sets and systems 118 (1), 21–45.

[29] Lee, M.-C., 2009. Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications 36 (8), 10896–10904.

[30] Lin, B., Wehkamp, R., Kanniainen, J., 2015. Practitioner's guide on the use of cloud computing in finance. Available at SSRN 2697583.

[31] Lin, J., Keogh, E., Lonardi, S., Chiu, B., 2003. A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM, pp. 2–11.

[32] Lin, X., Yang, Z., Song, Y., 2009. Short-term stock price prediction based on echo state networks. Expert systems with applications 36 (3), 7313–7317.

[33] Mittermayer, M.-A., 2004. Forecasting intraday stock price trends with text mining techniques. In: System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on. IEEE, pp. 10–pp.

[34] Nikfarjam, A., Emadzadeh, E., Muthaiyah, S., 2010. Text mining approaches for stock market prediction. In: Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on. Vol. 4. IEEE, pp. 256–260.

[35] Pai, P.-F., Lin, C.-S., 2005. A hybrid arima and support vector machines model in stock price forecasting. Omega 33 (6), 497–505.

[36] Ramachandran, M., Chang, V., 2014. Financial software as a service–a paradigm for risk modelling and analytics. International Journal of Organizational and Collective Intelligence 4 (3), 65–89.

[37] Shen, S., Jiang, H., Zhang, T., 2012. Stock market forecasting using machine learning algorithms. Sruthi. V is currently pursuing BE computer Science and Engineering in SSN College of Engineering Chennai, India. She is doing research in the field of machine learning.

[38] Ticknor, J. L., 2013. A bayesian regularized artificial neural network for stock market forecasting. Expert Systems with Applications 40 (14), 5501–5506.

[39] Vlachos, M., Kollios, G., Gunopulos, D., 2002. Discovering similar multidimensional trajectories. In: Data Engineering, 2002. Proceedings. 18th International Conference on. IEEE, pp. 673–684.

[40] Wang, J.-H., Leu, J.-Y., 1996. Stock market trend prediction using arima-based neural networks. In: Neural Networks, 1996., IEEE International Conference on. Vol. 4. IEEE, pp. 2160–2165.

[41] Winkler, W. E., 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage., 778–783.

[42] Zhang, Y., Glass, J. R., 2011. A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping. In: Interspeech. pp. 1909–1912.

[43] Zhang, Z., Jiang, J., Liu, X., Lau, R., Wang, H., Zhang, R., 2010. A real time hybrid pattern matching scheme for stock time series. In: Proceedings of the Twenty-First Australasian Conference on Database Technologies-Volume 104. Australian Computer Society, Inc., pp. 161–170.