Professor Kyoung-Sook Moon, PhD
E-mail: ksmoon@gachon.ac.kr
Mathematical Finance, Gachon University

Professor Hongjoong Kim, PhD
E-mail: hongjoong@korea.ac.kr
Department of Mathematics, Korea University

EFFICIENT ASSET ALLOCATION BASED ON PREDICTION WITH ADAPTIVE DATA SELECTION

Abstract. *Portfolio optimisation is a key issue in finance for individual investors and asset managers to make profits or hedge against market risks. Based on the analysis of financial time series using machine learning techniques with mathematically adaptive data selection, we predict the future trends of the stocks and make efficient profitable portfolio selection. It is the novelty of this work to express the training data as the union of subsets of similar data and build multiple machine learning networks, each one of which is specialised for the trends in one subset. Similar data in a subset gives ease of learning, which eventually leads to an improvement in prediction accuracy. The consideration of the possibility from one trend in the past to various outcomes in the future is another novelty. The portfolio management based on such improved learning gives the high rate of returns. When the proposed portfolio management scenario is applied to the stocks included in KOSPI index in Korea, up to 4 times more profits than the standard management are obtained.*

Keywords*: portfolio optimisation, stock market forecasting, deep learning, rebalancing scenario.*

JEL Classification: C15, C45, C63

1. Introduction

Allocation of assets to maximise expected returns and minimise costs such as risk is one of the most essential problems in finance. Traditionally, the Efficient Market Hypothesis (EMH) [1] implies that since asset prices fully reflect all the available information, it is hard to make a better prediction than the market. However, many studies report that the information from market data may predict asset fluctuations or yield returns better, [2, 3]. Also, it is being studied that superior results can be obtained by applying these predictions to portfolio optimisation models such as the existing mean-variance or omega models, [4, 5].

DOI: 10.24818/18423264/57.1.23.04

Recently, various machine learning (ML) techniques have been studied in many application fields including finance, to analyse vast amounts of data and make accurate predictions, [6, 7]. From single classifiers such as linear regression, decision tree, k-nearest neighbours and support vector machine to multiple classifiers such as random forests (RF) and majority voting, various ML methods are being used for financial market analysis, [8, 9]. Also, the deep learning (DL) algorithm, called *Long Short-Term Memory* (LSTM), is introduced by Hochreiter and Schmidhuber [10] and has been applied to achieve good results for financial market data [11, 12, 13].

But the accuracy of ML or DL prediction in the area of finance is not greater than those in other areas and is not always proportional to the amount, either. One of the reasons is that financial data shows oscillatory patterns due to high dimensional, nonstationary and complex nature of financial market, and thus similar trends of the financial data in the past may not lead to identical results in the future. Thus, the accuracy of the financial prediction depends not only on the quantity of the data, but also on its quality.

In order to compensate for these difficulties, we devise a new adaptive data selection method, and aim to improve accuracy by learning data similar to the patterns of recent trend so that the profit of the portfolio can be increased. Instead of constructing single ML or DL network which handles and learns all the financial patterns, we represent the training data by the union of subsets, each of which is a collection of data having similar patterns, and build multiple networks designed for the patterns in each one and only one subset. The distribution of tasks will lead to specialisation and eventually increase efficiency and accuracy of learning. Given a test data, the ML or DL network trained with the closest subset in terms of similarity is applied to its prediction. Here, the *Dynamic Time Warping* (DTW) method is used to measure the similarity between the data, [14, 15].

In order to choose the profitable assets, four different data construction methods are compared, including a method of collecting the data of similar trends based on DTW, called the *Adaptive Data Selection* (ADS-DTW), which we propose in this study. Then we perform *random forests* (RF), the ensemble learning method for classification, or LSTM to forecast up or down movement on the next rebalancing date. The results are analysed using the top 10 stocks in the KOSPI index in trading volume for 11 years from 2010 to 2020. The proposed ADS-DTW gives the best performance in return. The proposed method is generic, so that it can be combined with any machine learning methods.

The rest of the paper is organised as follows. Section 2 provides the summary of data and explains four data construction methods. Section 3 reviews the learning models, RF and LSTM, and portfolio selection scenarios. The experimental results with data analysis are presented in Section 4 and the conclusions are drawn in Section 5.

2. Data construction methods

2.1 Data

The stocks included in KOSPI index in Korea are used in the experiments. Top 10 stocks in KOSPI index in trading volume for 11 years from 2010 to 2020 are used in this study and the last 400 values are used as the test data, see Figure 1.

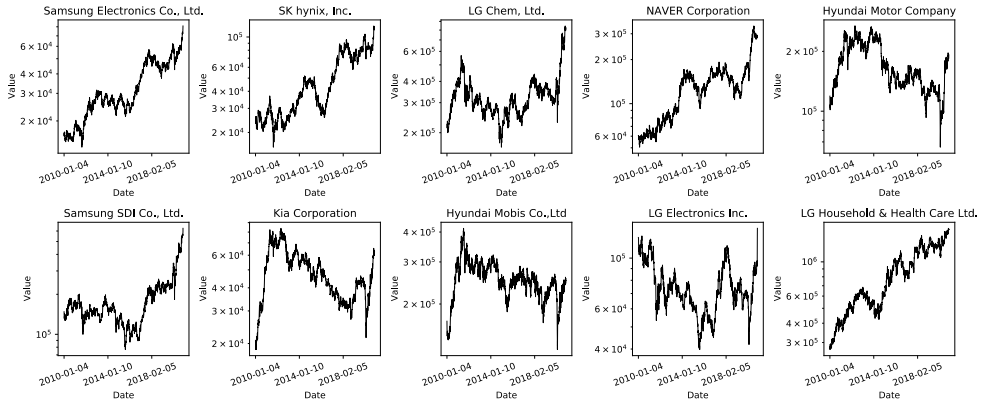

Figure 1. Financial data from top 10 stocks in KOSPI index in trading volume for 11 years from January, 2010 to December, 2020.

Table 1 describes the statistics including the count, mean, standard deviation, minimum, first-, second-, and third-quartiles, and maximum of the market data used.

Table 1. Statistics (count, mean, standard deviation, minimum, first-, second-, and third-quartiles, and maximum) of top 10 stocks in KOSPI index in trading volume for 11 years from January, 2010 to December, 2020.

Name	count	mean	std	min	25%	50%	75%	max
Samsung Electronics Co., Ltd.	2714	33157.41	13475.77	13600.00	23840.00	28200.00	45680.00	81000.00
SK hynix, Inc.	2714	48058.00	24106.55	15600.00	27600.00	39375.00	71900.00	120500.00
LG Chem, Ltd.	2714	332718.50	101727.39	164500.00	272500.00	313250.00	366500.00	846000.00
NAVER Corporation	2714	128640.78	55291.49	50538.00	80045.00	130185.00	157000.00	339000.00
Hyundai Motor Company	2714	169978.00	44349.82	65900.00	135000.00	158000.00	213500.00	268500.00
Samsung SDI Co., Ltd.	2714	179333.09	81170.55	76800.00	132500.00	156500.00	208500.00	628000.00
Kia Corporation	2714	48193.52	14420.18	18550.00	35912.50	46000.00	57975.00	83800.00
Hyundai Mobis Co.,Ltd	2714	254012.16	44323.75	129000.00	228500.00	251500.00	282000.00	414500.00
LG Electronics Inc.	2714	74672.54	17862.15	39800.00	62200.00	71600.00	85500.00	135000.00
LG Household & Health Care Ltd.	2714	825198.60	357053.03	270000.00	510000.00	786000.00	1153000.00	1649000.00

The financial data are oscillatory as in Figure 1, making it difficult to capture and learn its trend. The fluctuation in the data is smoothened out by using the exponentially weighted moving average in this study. Let us denote the asset

price $S_t \equiv S(t)$ at time t. Then the exponentially weighted moving average $\bar{S}_t$ is computed by

$$\begin{cases} \bar{S}_0 = S_0 \\ \bar{S}_t = (1-\alpha)\bar{S}_{t-1} + \alpha S_t \end{cases}$$

where $\alpha = 1 - \exp(-\ln(2)/halflife)$. The data used in Section 4 are smoothed out using $halflife = 1.75$.

After smoothing the data, the label is defined by the difference between the average of the past *w* smoothened data and the (non-smoothened) raw value in *p* days because the oscillatory pattern does not provide a reasonable label if the labelling is defined by the net difference of (non-smoothened) raw values between two days. The past *w* smoothened values after the standardisation procedure are provided as the feature values at each point. The parameters *w* and *p* are set to 20 and 10, respectively, in this study.

2.2 Training data construction

The training data is constructed in 4 different ways. The first approach is to use all the data except the test data as the training data, which will be called *Whole* in this study, and the other 3 methods define the training data with a subset of *Whole*. One way to consider a subset is to select the data that belong to a temporal period comparatively close to the test data, which will be called *Recent* below. In this study, data from the last 500 days are used as a training data set for *Recent*.

The next two methods introduce subsets of the training data and construct several ML networks, each of which is specialised for the patterns in each subset. One method uses *K-means* unsupervised clustering to partition the training data into several disjoint clusters. The *K-means* tries to partition the data into the groups of objects that are more related to each other than to objects in other groups. Then, a machine learning algorithm is applied to each cluster. For example, if the *K-means* splits the whole data into 5 clusters and if the LSTM method is considered, 5 LSTM-based neural networks are constructed, each of which is trained with one and only one cluster out of 5. Then, for each data in the test data, the *K-means* clustering is applied to identify the cluster it belongs to and the neural network trained with that specific cluster is used for its prediction. Such a method will be called *K-means*.

The other method to obtain subsets improves the *K-means* above in two aspects. Firstly, since unsupervised clustering may not be directly related with increase or decrease of stocks, we may gather mathematically similar data into the same subsets to form a cover of the training data. A *cover* of a set *X* in mathematics is a collection of its subsets whose union is the whole set *X*. The subsets will be constructed so that the data in the same subset of the cover have the same trends in terms of up-down movement. The method we propose measures the distance to identify the data of similar trends. Since the data we consider in this study are

financial time series, the *Dynamic Time Warping* (DTW) is used to measure the distance between two data. Contrary to L1 or L2 norms, DTW is efficient in measuring the difference between time series. The large DTW value implies that the patterns for two corresponding data are not close enough. See [14, 15] and references therein for more information on the DTW algorithms. Secondly, unsupervised clustering *partitions* the training data into disjoint subsets. That is, each data belongs to one and only one subset. But financial data shows oscillatory patterns and thus similar trends of the financial data in the past may not lead to identical patterns in the future. Thus, some trends should be considered for multiple outcomes or labels, which can be modeled by nonempty intersections of a cover. The resultant approach called Adaptive Data Selection with DTW (ADS-DTW) is explained in more detail below.

2.3 Adaptive data selection with DTW

Given the whole data available as the training data, a mathematical cover will be introduced. Let us denote the asset price $S_i \equiv S(t_i)$ at time t_i, $i = 1,2,\dots,N$ and the whole set $S = \{S_i, i = 1,2,\dots,N\}$. Consider covering of S, $S = \bigcup_{j=1}^{K} C_j$, where C_j, the subsets of S, are constructed as follows. The first step is to select one data, x_1, (the red circle in the Figure 2 (Left)) and find M nearest data (circles inside the blue circle in the Figure 2 (Left)) to define the first group, C_1. The distance is measured by DTW. The second step finds the farthest data from the center of the first group, called x_2, (the red circle in the Figure 2 (Middle)) and finds M nearest data (circles inside the green circle in the Figure 2 (Middle)) to define the second group, C_2. The intersection between C_1 and C_2 may not be an empty set. The next step is to find the farthest data from the centers of two previously defined groups, called x_3, (the red circle in the Figure 2 (Right)) and find M nearest data (circles inside the yellow circle in the Figure 2 (Right)) to define the third group, C_3. The intersection between two distinct C_j's may not be an empty set. The procedure is repeated until the entire data is covered.

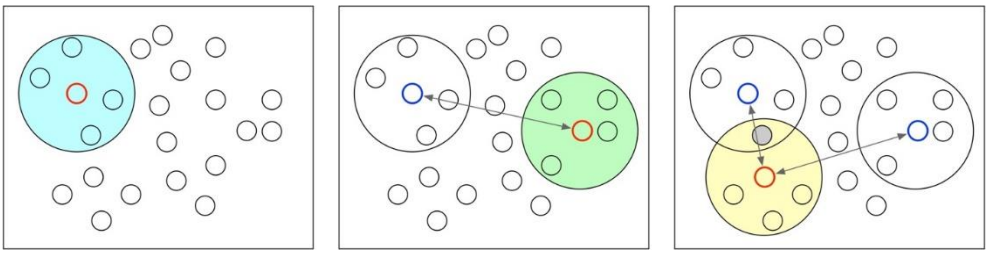

Figure 2. Construction of groups using ADS-DTW.

Then a machine learning algorithm is applied to each group similarly to the *K-means* approach in Section 2.2. For example, if the whole data is covered by 5 groups and the LSTM method is considered, 5 LSTM-based neural networks are constructed, each of which is trained with one and only one group. Then, for each data in the test data, the DTW-based distance between the given data and each centre of the groups is measured and the learning algorithm trained with the group corresponding to the closest distance will be used for the prediction of the given test data. *M* is set to 500 in this study.

Note that there may be an intersection between groups (like the filled circle in the Figure 2 (Right)), which is one of the differences from unsupervised clustering where the whole data are partitioned to disjoint clusters so that two different clusters have no elements in common. In case of the images of handwritten digits or animals, each image corresponds to one label, and partitioning from unsupervised clustering may be applicable. In case of the financial stock prices, on the other hand, similar trends may not lead to the same up or down movements. Thus, for such financial data, it is reasonable to educate the learning algorithm about the possibility of more than one outcome, and the consideration of non-empty intersection above may enhance the learning algorithm. The difference between partitioning with *K-means* and grouping with ADS-DTW is shown in the Figure 3 (Left). Note that in the partitioned data from *K-means* some data may be closer to the centre of other cluster and thus similarity may not be properly represented by the closest center. On the other hand, the groups from ADS-DTW are constructed based on the distance from the centres, so that similarity among the data within the same group is guaranteed. After dividing the data, we train the data of each group by applying machine learning techniques as shown in the Figure 3 (Right).

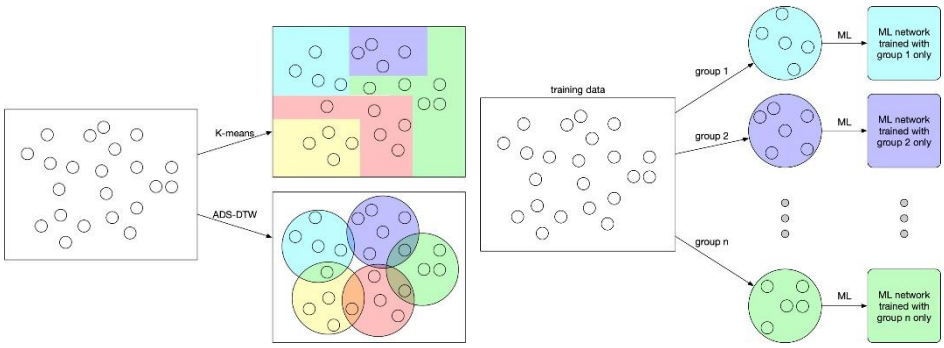

Figure 3. (Left) Comparison of groups between *K-means* and ADS-DTW and (Right) application of the machine learning method to each group.

3. Models

3.1. Learning models

First, let us take a brief look at the machine learning and deep learning methods we use. The *Decision Trees* are one of the easy-to-understand and well-used machine learning techniques as a method of classifying input data like a flow chart through questions in several stages. The *Random Forests* (RF) construct a multitude of these decision trees at training time and then ensemble their outputs. Random forests generally outperform decision trees and can be applicable to a wide range of problems, [16, 17].

Long Short-Term Memory (LSTM) is a kind of neural network that processes sequential data. By introducing a self-loop, a path through which the slope can flow for a long period of time is created, [10,18]. LSTMs have been found to learn long-term dependencies more easily than regular Recurrent Neural Networks (RNNs) and have been shown to perform well in financial time series processing problems, [11, 12, 13]. Figure 4 is a conceptual diagram of (Left) the RF and (Right) the LSTM.

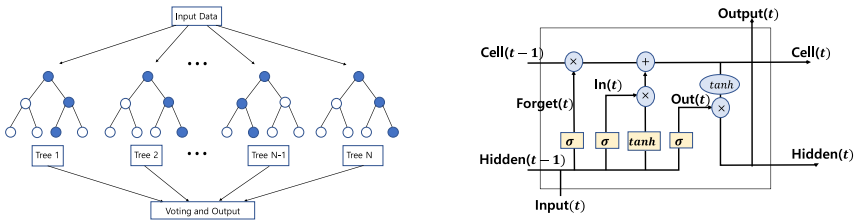

Figure 4. Conceptual diagram of (Left) the Random Forests and (Right) the Long Short-Term Memory

3.2. Portfolio selection scenarios

In the study, we construct a portfolio with 10 KOSPI stocks and run it for 400 days. We compare three portfolio management scenarios. The first scenario does not reorganise assets over the whole period once the assets are initially distributed. It is called *Buy & Hold* in this study. The other two scenarios restructure assets based on forecasts. The increase or decrease of each stock in 10 days is expected 40 times for 400 days, and one reallocates assets every 10 days when a new prediction is made. One scenario considers *Long only*. That is, the portfolio is constructed with stocks which are expected to increase in 10 days, and asset rebalancing is performed every 10 days, while no action is performed for the stocks which are expected to decrease. The other algorithm called *Long & Short* scenario additionally performs short selling. In other words, the portfolio consists

of the stocks which are expected to increase in 10 days similarly to the former, and short selling the stocks which are expected to decrease is additionally performed. The ADS algorithm for *Long & Short* portfolio management scenario is given as follows. The algorithm for *Long only* can be easily derived from this algorithm.

Algorithm 1: ADS Long & Short portfolio management algorithm

```
Data: Test data
Result: The value of the portfolio from Long & Short algorithm
construct the cover of the training set;
perform the ML training as in Figure 3 (Right);
while not at end of the time interval do
    evaluate the result of the previous portfolio construction;
    for each stock do
        find the closest group;
        predict using the ML algorithm trained with the closest group;
        if the stock is expected to decrease then
            short sell the stock;
        else
            add the stock to the portfolio;
        end
    end
    estimate the share of the stocks included in the portfolio;
end
evaluate the final value of the portfolio;
```

4. Empirical Experiments

4.1. Forecasting

We summarise the prediction accuracy for various data construction methods in Table 2. While RF seems to train properly, the LSTM seems to overfit slightly, especially in the *Whole* and *Recent* cases, and the result of the portfolio management below seems to be a consequence of this.

Figure 5 shows the ranges of prediction accuracies. The lower left bound of each bar represents the accuracies for the worst-case scenarios. The lower bounds of *Whole* and *Recent* go far below, while the bounds of *K-means* and ADS-DTW are relatively high, which shows that *K-means* and ADS-DTW are relatively reliable compared to *Whole* or *Recent*. That is, the construction of multiple networks specialised for the patterns in the subsets of the cover seems to be effective, leading to the excess of profit as shown in Section 4.2 below.

Table 2. Summary of the prediction accuracy using the training data construction methods, *Whole, Recent, K-means* and ADS-DTW.

		Whole	Recent	K-means	ADS-DTW	Mean
RF	Samsung Electronics Co., Ltd.	0.77	0.77	0.77	0.80	0.78
	SK hynix, Inc.	0.62	0.56	0.59	0.60	0.59
	LG Chem, Ltd.	0.75	0.77	0.70	0.70	0.73
	NAVER Corporation	0.74	0.58	0.66	0.74	0.68
	Hyundai Motor Company	0.65	0.66	0.66	0.61	0.65
	Samsung SDI Co., Ltd.	0.65	0.69	0.64	0.69	0.67
	Kia Corporation	0.81	0.74	0.81	0.80	0.79
	Hyundai Mobis Co.,Ltd	0.74	0.66	0.71	0.70	0.70
	LG Electronics Inc.	0.80	0.79	0.76	0.78	0.78
	LG Household & Health Care Ltd.	0.70	0.64	0.68	0.66	0.67
LSTM	Samsung Electronics Co., Ltd.	0.77	0.70	0.72	0.76	0.74
	SK hynix, Inc.	0.54	0.54	0.63	0.59	0.57
	LG Chem, Ltd.	0.72	0.74	0.72	0.77	0.74
	NAVER Corporation	0.62	0.68	0.68	0.70	0.67
	Hyundai Motor Company	0.67	0.64	0.62	0.64	0.64
	Samsung SDI Co., Ltd.	0.66	0.74	0.68	0.69	0.69
	Kia Corporation	0.71	0.74	0.72	0.73	0.72
	Hyundai Mobis Co.,Ltd	0.68	0.58	0.68	0.72	0.66
	LG Electronics Inc.	0.72	0.77	0.76	0.69	0.73
	LG Household & Health Care Ltd.	0.65	0.62	0.70	0.68	0.66
Mean	Samsung Electronics Co., Ltd.	0.77	0.73	0.74	0.78	0.76
	SK hynix, Inc.	0.58	0.55	0.61	0.60	0.58
	LG Chem, Ltd.	0.74	0.75	0.71	0.73	0.73
	NAVER Corporation	0.68	0.63	0.67	0.72	0.67
	Hyundai Motor Company	0.66	0.65	0.64	0.63	0.64
	Samsung SDI Co., Ltd.	0.66	0.72	0.66	0.69	0.68
	Kia Corporation	0.76	0.74	0.76	0.77	0.76
	Hyundai Mobis Co.,Ltd	0.71	0.62	0.69	0.71	0.68
	LG Electronics Inc.	0.76	0.78	0.76	0.73	0.76
	LG Household & Health Care Ltd.	0.68	0.63	0.69	0.67	0.67

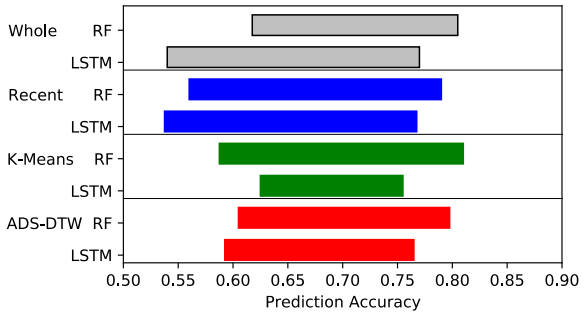

Figure 5. The ranges of prediction accuracies.

Table 3 shows the confusion matrices with respect to the training data construction method. Given the prediction for the upward or downward movement, the true positive (TP) and the true negative (TN) represent the numbers of correct predictions for the positives and negatives, respectively. Similarly, false positive (FP) and false negative (FN) show the numbers of wrong predictions for positives and negatives. The confusion matrix is a 2×2 matrix filled with TP, FP, FN, and TN. The accuracy (ACC) is a performance metric defined by

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN}$$

and the F_1-score is another metric defined by

$$F_1 = 2 \frac{PRE \times REC}{PRE + REC}$$

where the precision (PRE) and the recall (REC) are $\text{PRE} = \frac{TP}{TP+FP}$, $\text{REC} = \frac{TP}{TP+FN}$.

Table 3. The confusion matrices for (Top) RF and (Bottom) LSTM with respect to the construction method of the training data

RF

	Whole		Recent		K-means		ADS-DTW	
	up	down	up	down	up	down	up	down
up (prediction)	1889	416	1788	467	1810	437	1860	444
down (prediction)	694	1001	795	950	773	980	723	973
ACC	0.72		0.68		0.70		0.71	
F1-score	0.77		0.74		0.75		0.76	

LSTM

	Whole		Recent		K-means		ADS-DTW	
	up	down	up	down	up	down	up	down
up (prediction)	1757	479	1772	492	1821	481	1853	479
down (prediction)	826	938	811	925	762	936	730	938
ACC	0.67		0.67		0.69		0.70	
F1-score	0.73		0.73		0.75		0.75	

In Table 3 for RF, the prediction accuracy (ACC) and F1-score of *Recent* are relatively lower than those of the others, which leads to low profit of the portfolio in Section 4.2. In Table 3 for LSTM, the values of *Whole* and *Recent* are low, which affects the profit of the portfolio.

4.2. Portfolio Managements

Figure 6 (Top) shows the changes in the portfolio values when the prediction is made with RF and Figure 6 (Bottom) shows the changes with LSTM. Two left figures in Figure 6 are the results when only long is considered, while the two right figures are the results when short selling is also considered. In all four situations, ADS-DTW shows better performance than the others. Note that the results from *Whole* and *Recent* depend on machine learning methods. In particular,

it seems that the usage of the *whole* KOSPI stock data does not improve the prediction of LSTM.

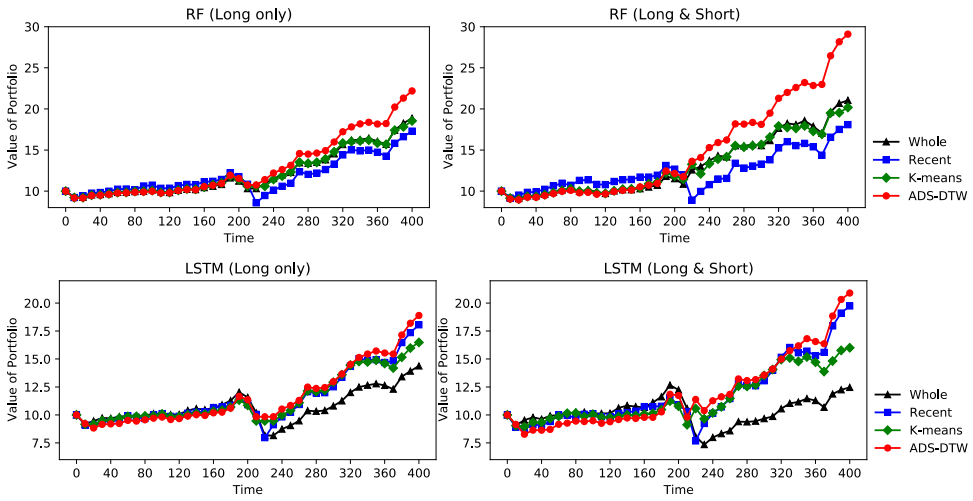

Figure 6. The changes of the portfolio values based on (Top) RF and (Bottom) LSTM and using (Left) 'long only' scenario and (Right) 'long & short' scenario.

In order to analyse Figure 6, let us compare the values of the portfolio with the prediction accuracy at each point in time. Figure 7 shows the *Long only* case with RF. Figure 7 (Top) represents the values of the portfolio from 4 training data construction methods and Figure 7 (Bottom) represents the prediction accuracy of the portfolio defined by Eqn. (1):

$$\frac{\text{(the number of stocks which are expected to increase and actually increase)}}{\text{(the number of stocks which are expected to increase)}} \quad (1)$$

at each rebalancing point. Note that ADS-DTW (red) shows better prediction accuracies between 280 and 360 so that its portfolio value increases more than the others. Such a superiority of the prediction accuracy seems to lead to the portfolio profit difference.

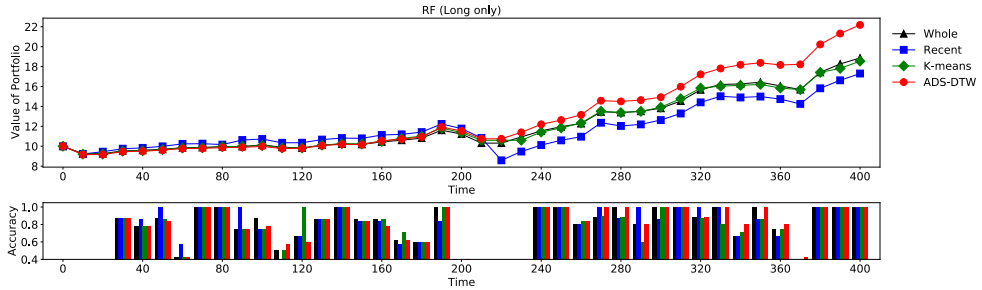

Figure 7. (Top) Comparison of the values of the portfolio between 4 training data construction methods and (Bottom) the prediction accuracy based on Eqn. (1) using the *Long only* scenario with RF.

Figure 8 shows the results based on the *Long & Short* policy and RF. Since long and short sell are both considered, the prediction accuracy at each point is defined by Eqn. (2):

$$\frac{\begin{array}{c}(\text{the number of stocks which are expected to increase and actually increase})\\ +(\text{the number of stocks which are expected to decrease and actually decrease})\end{array}}{(\text{the number of stocks})} \quad (2)$$

ADS-DTW (red in Figure 8) shows satisfactory predictions after 240 days at which its portfolio profit begins to outperform the others, while *Recent* (blue in Figure 8) shows poor predictions between 200 days and 240 days at which its portfolio begins to underperform the others. Note that the superiority of the prediction accuracy of DTW for *Long & Short* is better than that for *Long only*, which explains that the portfolio value from *Long & Short* is higher than the value from *Long only*. Similar analysis can be performed for LSTM (not shown) to support the advantage of ADS-DTW.

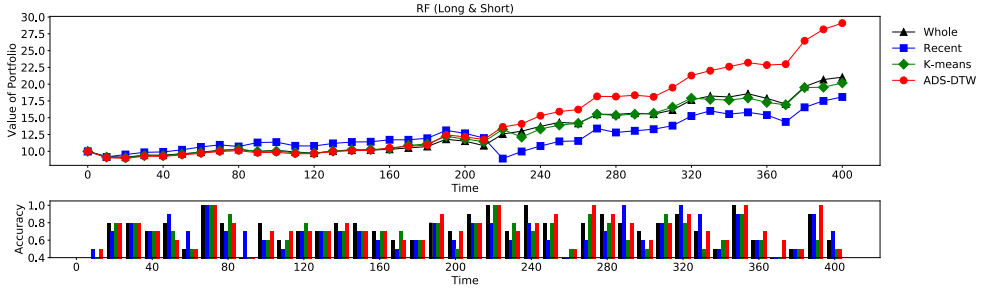

Figure 8. (Top) Comparison of the values of the portfolio between 4 training data construction methods and (Bottom) the prediction accuracy (Bottom) based on Eqn. (2) using the *Long & Short* policy with RF.

Figure 9 shows the excess of the profit from each portfolio management policy compared to the *Buy & Hold* as the benchmark. The advantage of ADS-DTW is confirmed again, and the excess of profit from ADS-DTW increases in time for both (Top) *Long only* policy and (Bottom) *Long & Short* policy. In case of *Long only* with RF, ADS-DTW results in almost double the profit of *Whole* (blue) and *K-means* (green) and three times the profit of *Recent* (orange) in **Figure 9**. In the case of *Long & Short* with RF, the advantage of ADS-DTW over the other three types gets bigger, and the profit from ADS-DTW exceeds that of *Recent* even more than 4 times. In case of LSTM (not shown), ADS-DTW still dominates the others in terms of the profit for both *Long only* and *Long & Short* cases.

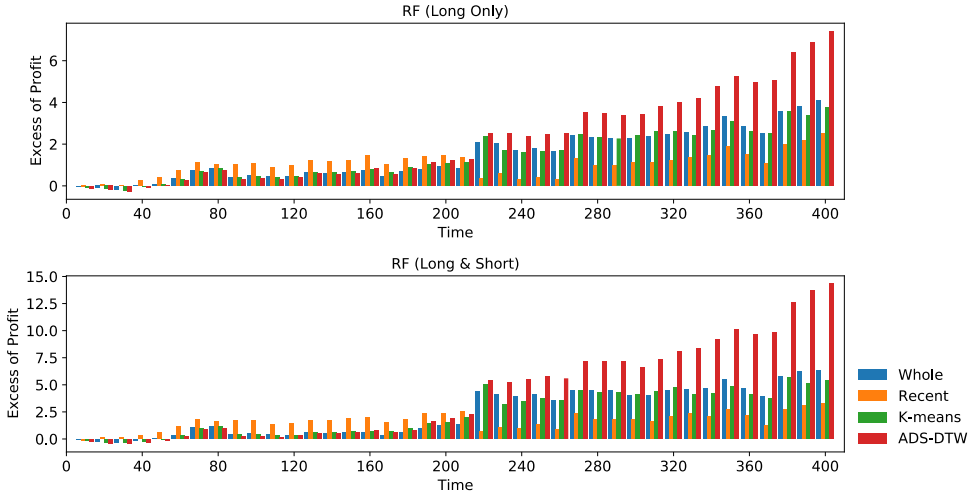

Figure 9. Comparison of the excess of the profit from (Top) *Long only* policy and (Bottom) *Long & Short* policy with RF.

Table 4 summarises the results of portfolio managements over a period of 400 days. For each machine learning method and the training data construction method, the value of the portfolio (value) at the end of the management, the profit from the initial capital (profit), and the ratio (ratio) using the result of *Buy & Hold* scenario, which initially selects stocks and holds over the whole period, as the benchmark are shown. In case of RF, the *Long only* scenario outperforms the *Buy & Hold* scenario and the *Long & Short scenario* outperforms both the *Long only* and *Buy & Hold* scenarios for all training data construction methods, and more importantly, ADS-DTW outperforms the other 3 methods. Note that ADS-DTW with the *Long & Short* scenario results in even 4 times the profit from the *Buy & Hold* scenario. In case of LSTM, the results are similar except that *Long & Short* with *Whole* is not better than *Buy & Hold*. But ADS-DTW still outperforms the

others and ADS-DTW with the *Long & Short* scenario makes 2.3 times the profit than the *Buy & Hold* scenario.

Table 4. Summary of the portfolio management using (Left) RF and (Right) LSTM for 4 training data construction methods and 3 policies.

	RF											
	Whole			Recent			K-means			ADS-DTW		
	value	profit	ratio	value	profit	ratio	value	profit	ratio	value	profit	ratio
Buy & Hold	14.77	4.77		14.77	4.77		14.77	4.77		14.77	4.77	
Long only	18.66	8.66	1.8	17.30	7.30	1.5	18.54	8.54	1.8	22.18	12.18	2.6
Long & Short	21.04	11.04	2.3	18.09	8.09	1.7	20.18	10.18	2.1	29.09	19.09	4.0

	LSTM											
	Whole			Recent			K-means			ADS-DTW		
	value	profit	ratio	value	profit	ratio	value	profit	ratio	value	profit	ratio
Buy & Hold	14.77	4.77		14.77	4.77		14.77	4.77		14.77	4.77	
Long only	14.38	4.38	0.9	18.06	8.06	1.7	16.48	6.48	1.4	18.89	8.89	1.9
Long & Short	12.46	2.46	0.5	19.74	9.74	2.0	16.00	6.00	1.3	20.90	10.90	2.3

The advantages of ADS-DTW are observed in the experiments, and those seem to result from two factors. One is the close similarity of trends in the subsets of the cover. Since they are quite similar, it is easy to be trained within an ML network and then to be specialised. The other factor is the flexibility from non-empty intersection of subsets, which can be led to possibly multiple outcomes in the financial markets. We are performing more on similarity analysis.

5. Conclusions

The current study proposes a quality data selection method which builds specialised networks and improves the value of the portfolio to enhance the profit. The method is validated by the stocks included in the KOSPI index in Korea.

For the analysis of the financial market, the amount of data makes it difficult to apply the machine learning or deep learning. Since ADS-DTW selects a portion of useful information, it may be applicable to data reduction so that other advanced learning methods such as reinforcement learning can be combined. ADS-like risk allocation to various assets using risk parity or risk budgeting approach and construction of a multichannel system for the prompt detection of several types of risk are on the future research agenda.

Acknowledgement

This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF-2018R1D1A1B07050046) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021R1F1A1054766).

REFERENCES

[1] **Fama, E.F. (1970), Efficient capital markets: A review of theory and empirical work.** *Journal of Finance*, 25, 383-417;

[2] **Malikiel, B.G. (2003), The Efficient Market Hypothesis and Its Critics.** *Journal of Economic Perspectives*, 17, 59-82;

[3] **Timmermann, A.G., Granger, C.W.J. (2004), Efficient market hypothesis and forecasting.** *International Journal of Forecasting*, 20, 15-27;

[4] **Ma, Y., Han, R., Wang, W. (2021), Portfolio optimization with return prediction using deep learning and machine learning.** *Expert Systems with Applications*, 165, https://doi.org/10.1016/j.eswa.2020.113973;

[5] **Yu, J.R., Chiou, W.P., Lee, W.Y., Lin, S.J. (2020), Portfolio models with return forecasting and transaction costs.** *International Review of Economics and Finance*, 66, 118-130;

[6] **Längkvist, M., Karlsson, L., Loutfi, A. (2014), A review of unsupervised feature learning and deep learning for time-series modeling.** *Pattern Recognition Letters*, 42, 11-24;

[7] **Henrique, B.M., Sobreiro, V.A., Kimura, H. (2019), Literature review: Machine learning techniques applied to financial market prediction.** *Expert Systems with Applications*, 124, 226-251;

[8] **Patel, J., Shah, S., Thakkar, P., Kotecha, K. (2015), Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques.** *Expert Systems with Applications*, 42, 259–268;

[9] **Moon, K.S., Jun, S., Kim, H. (2018), Speed up of the majority voting ensemble method for the prediction of stock price directions.** *Economic Computation and Economic Cybernetics Studies and Research*; ASE Publishing; 52(1), 215-228;

[10] **Hochreiter, S., Schmidhuber, J. (1997), Long short-term memory.** *Neural Computation*, 9(8), 1735-1780;

[11] **Fischer, T., Krauss, C. (2018), Deep learning with long short-term memory networks for financial market predictions.** *European Journal of Operational Research*, 270, 654-669;

[12] **Moon, K.S., Kim, H. (2019), Performance of Deep Learning in Prediction of Stock Market Volatility.** *Economic Computation and Economic Cybernetics Studies and Research*; ASE Publishing; 53(2), 77-92;

[13] **Song, D., Busogi, M., Chung Baek, A.M., Kim, N. (2020), Forecasting stock market index based on pattern driven Long Short-Term Memory.** *Economic Computation and Economic Cybernetics Studies and Research* 54(3), 25-41;

[14] **Keogh, E., Ratanamahatana, C.A. (2005), Exact indexing of dynamic time warping**. *Knowledge and Information Systems* 7, 358-386;

[15] **Kim, S., Lee, H., Ko, H., Jeong, S., Byun, H., Oh, K. (2018), Pattern Matching Trading System Based on the Dynamic Time Warping Algorithm. *Sustainability* 10(12), 4641:** https://doi.org/10.3390/su10124641;

[16] **Géron, A. (2019), *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, *O'Reilly Media*,** ISBN 978-1492032649**;**

[17] **Krauss, C., Do, X.A., Huck, N. (2017), Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500.** *European Journal of Operational Research* 259(2), 689-702;

[18] **Goodfellow, I., Bengio, Y., Courville, A. (2016), *Deep learning (Adaptive Computation and Machine Learning series)*,** *The MIT Press*, ISBN 978-0262035613.