16th International Conference on Innovative Computing, Information and Control (ICICIC2022)
Onsite (Chongqing, China) and Online, September 15-16, 2022

Ensemble method for feature selection in time-series prediction based on image similarity

Hyunjae Lee[1], Dohee Kim[2], Sangjae Lee[2], Hanbyeol Park[2],
Hyerim Bae[2,*], and Keonwoo Choi[3]

[1] Department of Industrial Engineering, Pusan National University, Busan 46241, Korea;
hyunjea414@pusan.ac.kr (H.L.);

[2] Major of Industrial Data Science and Engineering, Department of Industrial Engineering, Pusan National University,Busan 46241, Korea; kimdohee@pusan.ac.kr (D.K.); selflsj@pusan.ac.kr (S.L.);

pb104@pusan.ac.kr (H.P.); hrbae@pusan.ac.kr (H.B.);

[3] Shipping Big Data Analysis Center, Korea Maritime Institute, Busan, 49111, Korea;
ak8102@kmi.re.kr(K.C.)

* Correspondence: hrbae@pusan.ac.kr; Tel.: +82-51-510-2733

Received XXX 2022; accepted XXX 2022

ABSTRACT. This paper proposes the implementation of an ensemble feature selection method based on image similarity in time-series data. It is important to find the optimal lag of a feature while analyzing time-series data as the presence of lags in time-series data increases significantly the number of features to be considered. Feature selection plays an important role in distinguishing which features are important and finding features that have a meaningful effect on improving prediction performance. Feature selection based on only numerical data has limitations with respect to predicting rapidly changing time-series data. In this work, we present the availability of our ensemble feature selection method where multiple subsets from feature selection methods based on image similarity are combined into a single subset. We show that our proposed method exhibits better performance than a single feature selection technique based on only image similarity or number.

Keywords: Ensemble, Feature selection, Time-series, Lag, Image similarity

1. Introduction. Preemptive decision-making based on accurate predictions of the future is critical for shipping companies[1]. Appropriate decisions using timely and accurate container freight forecasting can result in significantly reduced costs for shippers and carriers[2]. Owing to the COVID-19 pandemic, the container freight index has risen significantly, making it more difficult to predict[3]. Therefore, it is necessary to develop an accurate and reliable forecasting container freight index model. For this reason, this study focuses on the Shanghai Containerized Freight Index (SCFI). We analyzed the data to find relevant features for more accurate SCFI prediction.

To analyze time-series data, a lag should be considered for each feature. If the lag is not accurately identified, features with important relationships cannot be identified[4]. However, a disadvantage when considering lags is that the number of features to be considered increases, and it is necessary to determine which lag can help predict the performance of a specific feature.

Owing to redundant features and the curse of dimensionality, the prediction performance of time-series models can decrease in a high-dimensional dataset. To solve this problem, we present a feature-selection method. Feature selection is the process of identifying and selecting the relevant features from the original dataset so that a model can

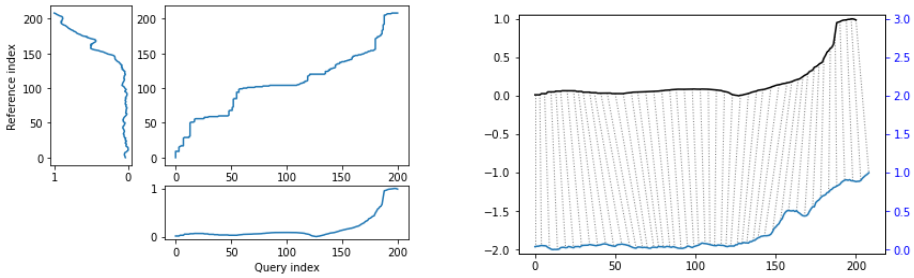
focus on distinguishing features that are useful for prediction[5]. Feature selection has three advantages: (i) improvement of prediction performance, (ii) selection of relevant features, and (iii) reduction of time required for analysis and learning[6]. Typically, the lag with the highest cross-correlation efficiency is selected as the optimal lag between a target and its features[7]. However, in real-world situations, temporal variations in data do not represent simple regularities, making them difficult to analyze and predict accurately. Only correlations and their combinations do not explain the characteristics of real data[8]. The convolutional neural network (CNN) model filters out noise from image-based data and extracts more relevant features, allowing learning algorithms to focus on important features and perform better future predictions [9]. Graph shape-based feature selection methods do not select features only with numerical data values but find related features even if they rise rapidly through trends or have different units. We propose a feature selection method that ensembles mean square error (MSE), dynamic time distortion (DTW), cosine similarity, and img2vec based on image similarity. Finally, we compared with each single feature selection subset to verify its performance.

2. Background. In this section, we describe the algorithm used in our ensemble method in this study. It contains a definition and overview of the MSE, DTW, cosine similarity, and correlation.

2.1. Mean Squared Error (MSE). The MSE is widely used to assess image quality and distortion because it is easy to calculate and does not require significant computational complexity[10]. The MSE is also a full reference of the image pixels and is averaged by the image size. The closer the value is to zero, the greater the similarity between two images. Sara et.al.(2019) defined the MSE between two images, where g(x, y) is a feature and $\hat{g}(\mathrm{x}, \mathrm{y})$ is the target[11].

$$\mathrm{MSE} = \frac{1}{MN}\sum_{x=0}^{M}\sum_{y=1}^{N}[\hat{g}(x,y) - g(x,y)]^2 \quad (1)$$

2.2. Dynamic Time Warping (DTW). Research on the indexing of time series is mainly conducted by Euclidean distance measurement method. DTW, one of the methods of indexing time series, is a stronger distance measure for time series by matching similar shapes between time series and is widely used in various fields including science, medicine, and finance[12]. Some studies have demonstrated that DTW is better for indexing than the Euclidean distance[13, 14]. The advantage of DTW is that it can be applied to time series with different time sequences. DTW represents the distance based on the path, which means that the smaller the distance, the more similar the pattern[12]. We utilized these DTW characteristics for feature selection to select relevant features with a small distance between the SCFI and features. Figure 1 shows an example of the application of DTW.



(a) (b)

FIGURE 1. An example of DTW
(a) A warping matrix to search for an optimal path to align the time sequences (b) The alignment result

2.3 Cosine Similarity. Cosine similarity is a simple and effective way to measure the angle between two normalized vectors as a dot product. The result of the cosine similarity measure is always within the range of -1 and +1. The closer it is to +1, the more positive is the similarity between the two vectors. Baoli Li et.al.(2013) defined cosine similarity between two vectors X and Y in the N dimension [15, 16].

$$\text{Cosine Similarity}(X, Y) = \frac{X \bullet Y}{\|X\| \, \|Y\|} = \frac{\sum_{i=1}^{N} X_i \times Y_i}{\sqrt{\sum_{i=1}^{N} {X_i}^2} \sqrt{\sum_{i=1}^{N} {Y_i}^2}} \tag{2}$$

img2vec model provided by Keras was used as a preprocessing method to obtain a vector. This library uses the ResNet50 model, which is pre-trained on ImageNet, to generate image vectors[17]. We measure the image cosine similarity between the generated image vectors.

2.4 Pearson Correlation Coefficient. The Pearson correlation coefficient is a metric of the linear correlation between two variables X and Y in statistics, and it is widely used in various applications to analyze linear correlations. The covariance of the two variables was divided by the product of each standard deviation and digitized as a value between -1 and +1. The value +1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no linear correlation. The equation for the Pearson correlation coefficient is as follows[18]:

$$\text{Pearson Correlation Coefficient}(X, Y) = \frac{\sum_{i}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i}^{n} (X_i - \bar{X})^2 \, (Y_i - \bar{Y})^2}} \tag{3}$$

3. Proposed method. This section describes our proposed method based on the ensemble method. The ensemble is contained using the image similarity measurement method described in Section 2. In the feature-selection method based on an ensemble, multiple feature subsets considering lags from each feature selection are combined to create an optimal subset of features with an optimal lag that improves prediction performance[7].

Our proposed method consists of two steps for creating an optimal subset of features with an optimal lag. The first step involves creating each subset using each feature-selection method. Then, multiple subsets are aggregated into the last subset. Each subset is created according to the feature-selection methods presented previously based on the criterion for

each method in the state of considering lag of 8-weeks ago, 4-weeks ago, no lag, 4-weeks later, and 8-weeks later. From the perspective of cosine similarity, we only selected features with a cosine similarity higher than 0.8, which represents a similar vector direction. Among the features considering all lags, the lag that was subsequently presented was over 0.8 when compared to the target graph vector, and it had the highest value, and the subsets were created using the selected features. Consistent with the previous method, feature selection using the DTW and MSE also needs to set a criterion. As mentioned earlier, DTW and MSE are similar because the values are smaller. We selected features within the top 20%. If there are different lags for one feature, the lag with the smallest value for each feature was finally selected.

In the second step, we focused as much as possible on using the features of the subsets derived from each method. First, we used all features selected in each method. If the same feature is selected in the three image similarity measurement methods and two lags out of the three are the same, the lag with more is selected. However, in this case, if all the lags are different, the lag with the highest positive correlation coefficient is selected. Figure 2 describes the framework of our ensemble method.

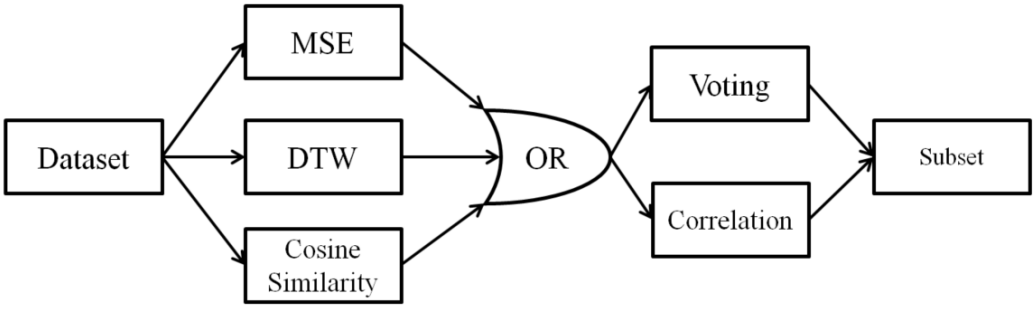


FIGURE 2. A framework of the ensemble method

4. Experiment. In this section, we present a description of our dataset and the results of our proposed method based on ensemble feature selection. The prediction experiment performed predictions at 1-week, 4-weeks, 12-weeks, and 24-weeks. We analyzed whether our proposed method is suitable for the short-term, mid-, and long-term predictions of our data.

4.1 Dataset. The dataset was provided by the Korea Maritime Institute (KMI), which focuses on research and development in the shipping sector. In Table 1, 18 variables, including the SCFI, External Variables, and Port Congestion Index, were used as data. The target to be predicted is the SCFI, and the feature used to predict it has a total of 17 variables, excluding the target. The data used to build the model were from January 1, 2018, to December 31, 2021. The number of data points used was 209 (weekly).

TABLE 1. Description of dataset

	Group name	Variable name	Unit
Target	SCFI	SCFI (Comprehensive)	Index
Feature	External Variables	Average Earnings	$/day
		Bunker Price	$/Tone
		Total Container ships Number	No
		Total Container ships TEU	TEU
		Newbuilding Prices	$m

		(1650/1850 TEU)	
		Newbuilding Prices (13000/14000 TEU)	$m
		Newbuilding Prices (3500/4000 TEU)	$m
		Newbuilding Prices (13000/13500 TEU)	$m
		5 Year Finance based on Libor	$m
	Port Congestion Index	Port congestion index (Comprehensive)	Million TEU
		Port congestion index - East coast North America	Million TEU
		Port congestion index - West coast North America	Million TEU
		Port congestion index - The United Kingdom	Million TEU
		Port congestion index - Mediterranean/Black Sea	Million TEU
		Port congestion index - East Asia	Million TEU
		Port congestion index - South East Asia	Million TEU
		Port congestion index - China	Million TEU

4.2 Result of subsets configured through each feature selection.

- MSE and DTW: Feature selection based on MSE selects features with an image similarity value in the top 20%, and feature selection based on DTW selects features according to the DTW distance in the top 20%. The top 20% of features were used to determine the correlation and cosine similarity. Four features were selected according to each lag as a result of factors with a value of 633.128, corresponding to a threshold of 20% in MSE. Four features according to each lag were selected as a result of factors for which there is a value of 10.44 as a threshold in DTW.

- img2vec & Cosine Similarity: After extracting the feature vector of the feature graph image considering all lags using the pre-trained img2vec model, we obtain the cosine similarity with the feature vector of the target graph image. Subsequently, we selected features with cosine similarity values greater than 0.8. If several lags are selected for one feature, the lag with the highest cosine similarity value is selected. Subsequently, nine features are selected. Figure 3 shows a two-dimensional (2D visualization of the clustering for each graph image using t-SNE. It can be confirmed that images with similar feature vectors were similar to each other.

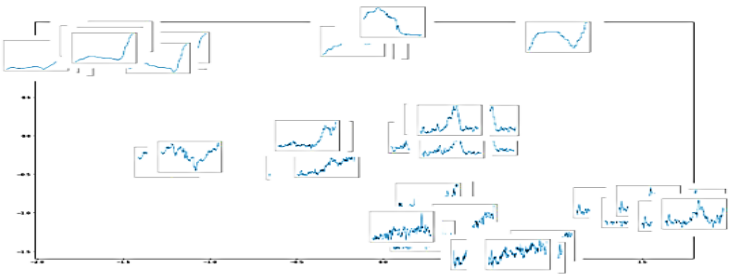


FIGURE 3. t-SNE visualization.

- Correlation: The correlation coefficient between the target SCFI and features was analyzed, considering all lags. Then, we selected a value over 0.8, which is generally considered a highly relevant correlation. For one specific feature, the lag with the highest correlation coefficient value was selected. In this way, a total of seven features were selected.
- Proposed method: As shown in Table 2, eleven variables were selected for the ensemble feature selection presented in Section 3.

TABLE 2. A subset of feature selection.

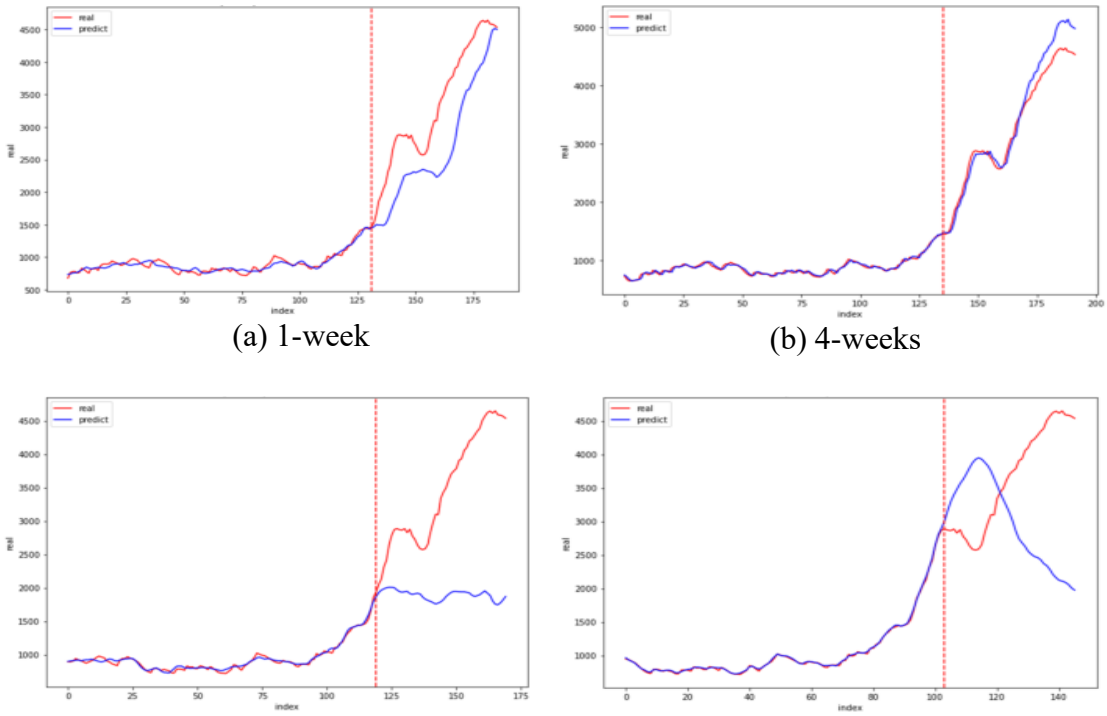
Group name	Variable name	MSE	DTW	Img2vec & Cosine Similarity	Correlation	**Proposed Method**
SCFI	SCFI (Comprehensive)	√	√	√	√	**√**
External Variables	Average Earnings	√ (-4 week lag)	√ (+8 week lag)	√ (-8 week lag)	√ (-8 week lag)	**√ (-8 week lag)**
	Bunker Price					
	Total Container ships Number			√ (+8 week lag)		**√ (+8 week lag)**
	Total Container ships TEU			√ (+4 week lag)		**√ (+4 week lag)**
	Newbuilding Prices (1650/1850 TEU)			√ (+4 week lag)		**√ (+4 week lag)**
	Newbuilding Prices (13000/14000 TEU)	√ (-8 week lag)		√ (Original)	√ (-8 week lag)	**√ (-8 week lag)**
	Newbuilding Prices (3500/4000 TEU)	√ (+8 week lag)	√ (Original)	√ (+8 week lag)	√ (-8 week lag)	**√ (+8 week lag)**
	Newbuilding Prices (13000/13500 TEU)	√ (-8 week lag)	√ (-8 week lag)	√ (+8 week lag)	√ (-8 week lag)	**√ (-8 week lag)**
	5 Year Finance based on Libor			√ (-4 week lag)		**√ (-4 week lag)**
Port Congestion Index	Port congestion index (Comprehensive)		√ (+8 week lag)			**√ (+8 week lag)**
	Port congestion index - East coast North America				√ (-8 week lag)	

	Port congestion index - West coast North America			√ (-8 week lag)	√ (+4 week lag)	**√ (-8 week lag)**
	Port congestion index -United Kingdom				√ (-4 week lag)	
	Port congestion index - Mediterranean/Black Sea					
	Port congestion index - East Asia					
	Port congestion index - South East Asia					
	Port congestion index - China					
Total number of features used		5	5	10	8	**11**

4.3 Result of the experiment. The model that was used in the experiment is the LSTM model, which is widely used for predicting time-series data. The RMSE and MAPE were used to evaluate the prediction performance. The results of the experiment are shown in Table 3, and the plots showing the results of the proposed method are shown in Figure 4.

TABLE 3. Result of the experiment.

Prediction period	Subset	2018.01.01 ~ 2021.12.31	
		RMSE	MAPE
1 week	MSE	1166.705	30.822
	DTW	1263.25	32.489
	Cosine similarity	571.524	14.978
	Correlation	1225.578	27.745
	Proposed method	**280.682**	**7.86**
4 week	MSE	832.383	18.824
	DTW	581.432	13.729
	Cosine similarity	750.545	16.109
	Correlation	527.881	12.964
	Proposed method	**509.672**	**13.276**
12 week	MSE	2264.018	53.469
	DTW	2134.867	54.278
	Cosine similarity	2194.517	55.624
	Correlation	2704.516	59.958
	Proposed method	**1876.258**	**45.21**
24 week	MSE	2483.673	62.771
	DTW	1926.935	49.039
	Cosine similarity	**1236.007**	**30.869**
	Correlation	2109.019	53.692
	Proposed method	1243.406	31.145



(a) 1-week (b) 4-weeks

(c) 12-weeks (d) 24-weeks

FIGURE 4. Graph visualization of prediction results each period

5. Conclusion. In this study, we propose ensemble method for feature selection based on image similarity in rapidly changing time-series data. We introduce the idea of feature selection based on image similarity to find relevant features that can improve the prediction performance, even though lags are applied. We experimented short-term and mid-long-term predictions using the Shanghai Containerized Freight Index(SCFI), which has risen significantly to the COVID-19 pandemic. We showed that our proposed method outperformed other single-feature selection methods with respect to short-term predictions corresponding to the 1-week and 4-week predictions. In addition, the mid-term prediction corresponding to the 12-week prediction was better than the other methods, but further analysis of the graph confirmed that the prediction graph did not follow the actual graph well. Our proposed method was not the best method for making a long-term prediction corresponding to the 24-week prediction, but it was the next best. All methods, as well as the proposed method, show poor performance in mid- to long-term predictions. This is because the amount of data for mid- to long-term predictions was not sufficient, so learning was not done well. After collecting more data, we plan to ensure that the proposed method outperforms the single feature -selection method even in mid- to long-term predictions.

In conclusion, we propose a method to show good prediction performance overall. Various image-based feature selection methods are combined to sufficiently reflect external influences that are not reflected in each feature selection method, while reducing dimensions, preventing data duplication, and increasing suitability. Through this, it was possible to provide explanatory power and better predictive performance for the influence of each shape. In future works, we will also verify that our proposed method performs well when applied to other types of time series data.

Acknowledgment. This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2020R1A2C1102294) and supported by KMI (the Korea Maritime Institute), NRC(the National Research Council for Economics, Humanities and Social Sciences)

REFERENCES

[1] Stopford, M. Maritime Economics 3e; Routledge: London, UK, 2008.

[2] Munim, Z.H., Schramm, HJ, Forecasting container shipping freight rates for the Far East – Northern Europe trade lane, *Marit Econ Logist* vol.19, no.1, pp.106–125, 2017.

[3] ARDELEAN, Adina, et al, Fathoming Shipping Costs, 2022.

[4] Olden, Julian D., and Bryan D. Neff, "Cross-correlation bias in lag analysis of aquatic time series.", *Marine Biology* 138.5 pp.1063-1070, 2001.

[5] Hoque, Nazrul, Mihir Singh, and Dhruba K. Bhattacharyya, "EFS-MI: an ensemble feature selection method for classification.", *Complex & Intelligent Systems* vol.4, no.2, pp.105-118, 2018.

[6] Rodríguez, Daniel, et al. Detecting fault modules applying feature selection to classifiers, *IEEE International Conference on Information Reuse and Integration*, pp.667-672, 2007.

[7] Wei WWS, Time Series Analysis: Univariate and Multivariate Methods. Addison-Wesely, New York, 1990.

[8] Kanad Chakraborty, Kishan Mehrotra, Chilukuri K. Mohan, Sanjay Ranka, Forecasting the Behavior of Multivariate Time Series Using Neural Networks, Neural Netw, vol.5, no.6, pp.961-970, 1992.

[9] Livieris, I.E., Pintelas, E.G., & Pintelas, P.E, A CNN–LSTM Model for Gold Price Time-series Forecasting. *Neural Comput. and Appl*, pp.1-10, 2020

[10] Zhou Wang and A. C. Bovik, A Universal Image Quality Index, in IEEE Signal Process. Lett., vol. 9, no. 3, pp. 81-84, 2022

[11] Sara, U. , Akter, M. and Uddin, M, Image Quality Assessment through FSIM, SSIM, MSE and PSNR, A Comparative Study. J. of Comput. and Commun., vol.7, no.3, pp.8-18, 2019.

[12] Keogh, Eamonn, and Chotirat Ann Ratanamahatana, Exact indexing of dynamic time warping, *Knowledge and information systems* vol.7, no.3, pp.358-386, 2005.

[13] Bar-Joseph, Ziv, et al, A new approach to analyzing gene expression time series data, *Proceedings of the sixth annual international conference on Computational biology*, 2002.

[14] Aach, John, and George M. Church, Aligning gene expression time series with time warping algorithms, *Bioinformatics*, vol.17, no.6, pp.495-508., 2001.

[15] Nguyen, Hieu V. and Li Bai, Cosine Similarity Metric Learning for Face Verification, *ACCV, 2010* .

[16] LI, Baoli; HAN, Liping, Distance Weighted Cosine Similarity Measure for Text Classification, In: Int. Conf. on Intell. Data Eng. and Autom. Learning. Springer, Berlin, Heidelberg, pp. 611-618, 2013.

[17] Jaredwinick, Image to Dense Vector Embedding, 2019. Available online: https://github.com/jaredwinick/img2vec-keras (accessed on 20 June 2022).

[18] Benesty, Jacob; Chen, Jingdong; Huang, Yiteng, On the Importance of the Pearson Correlation Coefficient in Noise Reduction, IEEE Trans. On Audio, Speech, and Lang,

Process, vol.16, no.4, pp.757-765, 2008.